Crystallization and Structure Determination of Coronavirus Main Proteinases

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

Vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät der

Friedrich-Schiller-Universität Jena

Von MSc Kanchan Anand aus Delhi (Indien)

Jena 2002

Acknowledgments

I would like to express my sincere gratitude to Prof. Hilgenfeld for giving me the opportunity to understand protein crystallography and providing all the facilities promptly to pursue my work. His stimulating and lively discussions have always been very helpful. I especially acknowledge the guidance and insight he imparted to work in the scientific community and the freedom he granted in pursuing this work.

Special thanks to Dr. John Ziebuhr (University of Würzburg) for preparation, purification, characterization of TGEV M^{pro}, HCoV M^{pro}, their mutant proteins, many helpful discussions and for his generous support.

Thanks to expert beamline staff at EMBL (DESY Hamburg) and at ELETTRA (Trieste, Italy) for their assistance in data collection. I gratefully acknowledge Dr. Gührs (IMB, Jena) and Dr. N. Oldham (Max Planck Institute for Chemical Ecology, Jena) for MALDI-TOF experiments.

My friends and colleagues at the Institut für Molekulare Biotechnologie, past and present, deserve thanks for their generous help and for creating a warm atmosphere: Bartholomeus Kuettner, Tom Sicker, Jeroen Mesters, Debnath Pal, Tanis Hogg, Gottfried Palm, Manfred Weiss, Guido Kaulmann, Ursula Kaulmann, Alan Tunicliffe, Santosh Panjikar, Andre Vogel, Dietmar Schwertner, Silke Schulz, Astrid Rau, Frau Härtl, Thomas Klupsch, Peter Mühlig, Axel Walter, Sundaram, Aida, Daniella and administrative staff at IMB.

I am immensely thankful to Mr. & Mrs. Budach, Trevor & Efi Fenning, Mr. & Mrs. Reinke, Mr. & Mrs. Baljinder Singh and Parvesh, for they were always there for me.

Thanks to everyone at Delhi who have made me what I am today! Specially to Nanhe, Babbu bhaiya, Suman, Madhu, Rajni, Sushamadi, Guddidi, Ayushi, Dr. Upadhyaya, Dr. Dixit, IP Singh, Ritu, Maneesha, Amita, Anita, Prasad, also to friends in Jena, Dr. Nasheuer, Dr. Ramachandran & Kalyani, Cyril, Renata, Pavlina, Frank, Andre, Heather, Oxsana, Dr. Malvia, Dr. Shantanu & Puja Rastogi, Zafar, Manju, Kamran, Dr. Farooq, Chandrashekhar, Altaf, Rahul, Mona, Pulkit, and Maria Angela for their great company.

Last but not the least, sincere thanks to my parents back home in Delhi (India); words alone cannot express how much your love, undiluted support and encouragement have inspired me. ***This work is dedicated to you***.

TABLE OF CONTENTS

Page No.

FIGURE INDEX

TABLE INDEX

ABBREVIATIONS

AA/aa	amino acid
Antipain	[(S)-I-Carboxy-2-Phenylethyl]-Carbamoyl-L-Arg-L-Val-Arginal ($C_{27}H_{44}N_{10}O_6$)
ASA	accessible surface area ($Å^2$)
a.u.	asymmetric unit
B-Factor	Temperature factor ($Å^2$)
BSA	Bovine serum albumin
DLS	Dynamic Light Scattering
DTT	1,4-Dithiothreitol
E. coli	*Escherichia coli*
$\|F_{hkl}\|$	the absolute value of the structure factor amplitude
HEPES	2-[4-(2-Hydroxyethyl)-1-piperazinyl]ethanesulphonic acid
I	reflection intensity
k	a scale factor
kDa	kiloDalton
M	moles/litre
MAD	multiple wavelength anomalous dispersion
ml	millilitre
mM	millimolar
MPD	(±)2-Methyl-2,-4 pentanediol
$M^{pro(s)}$	main proteinase(s)
mRNA	messenger RNA
µl	microlitre
MALDI- TOF	Matrix assisted laser desorption ionization time of flight
NCS	non-crystallographic symmetry
PAGE	polyacrylamide gel electrophoresis
Pefabloc SC	[4-(2-Aminoethyl)benzenesulfonyl fluoride·HCl] - serine protease inhibitor ($C_8H_{10}NSO_2F \cdot HCl$)
PEG 6000	polyethylene glycol 6000

PDB	Protein Data Bank
rmsd	root-mean-square deviation
RNA	ribonucleic acid
S	Svedberg unit
SDS-PAGE	sodium dodecyl sulfate gel polyacrylamide electrophoresis
(SGPA)	*Streptomyces griseus* proteinase A
TLCK	Nα-p-Tosyl-$_L$-lysine-chloromethyl ketone ($C_{14}H_{21}ClN_2O_3S.HCl$)
TPCK	Nα-p-Tosyl-$_L$-phenylalanine-chloromethyl ketone ($C_{17}H_{18}ClNO_3S$)
v/v	volume/volume
V_m	Matthews coefficient (**Vm = V/(n*m)**, $Å^3$/dalton
w/v	Weight/volume

1. INTRODUCTION

Transmissible gasteroenteritis virus (TGEV) and human coronavirus (HCoV) belong to the *Coronaviridae* family which has two genera: coronavirus and torovirus. Together with the *Arteriviridae*, the *Coronaviridae* form the order *Nidovirales* (latin *nidus*, nest) and produce mRNAs in an extensive nested-set arrangement.

TGEV (strain Purdue)[§] and HCoV (strain 229E)[§] are group-1 coronaviruses that are most closely related to feline infectious peritonitis virus (FIPV), avian infectious bronchitis virus (IBV), and murine hepatitis virus (MHV), representing the prototypes of the coronavirus groups III and II, respectively.

1.1 Infections / Diseases

Coronaviruses cause common respiratory and enteric diseases in humans and domestic animals (Myint, 1995; Johnston & Holgate 1996). HCoV is one of the main causes of upper respiratory tract infections ('common colds') in humans, but also lower respiratory tract illness and gastroenteritis have been reported (Zhang et al., 1994; Myint, 1995). TGEV is mainly associated with profuse, watery gastroenteritis in piglets, which frequently causes severe dehydration and death within 2 to 5 days after infection (for reviews, see Enjuanes & van der Zeijst, 1995; Saif & Wesley, 1999). Other members of this family are IBV, the first coronavirus to be isolated from the domestic fowl, bovine coronavirus (BcoV), turkey coronavirus (TCV), FIPV, canine coronavirus (CCV) and porcine epidemic diarrhoea virus (PEDV). MHV and porcine haemagglutinating encephalomyelitis virus (HEV) are well-known causative agents of neurological diseases.

1.1.1 Viral host ranges

Coronaviruses have relatively restricted host ranges, infecting only their natural hosts and closely related animal species. Occasionally, cross-species infection by coronaviruses occurs. Interestingly, the host range of coronaviruses is associated with the receptor usage.

[§] All the studies in this work are based on these strains.

1.1.2 Virus attachment

The first step in viral infection is the binding of the virus to target cells. Several coronaviruses including IBV, BCoV, and some strains of MHV and HCoV, can cause hemagglutination (Sugiyama & Amano, 1980; Schultze *et al.*, 1990; Zhang *et al.*, 1994). The binding residue on the cell surface is a 9-O-acetylated neuraminic acid moiety of glycoproteins or glycolipids (Schultze *et al*., 1990), although different coronaviruses may prefer different structural isoforms of 9-O-acetylated neuraminic acid.

1.1.3 Receptors

For the establishment of viral infection, more specific binding between virus and cell is required for the coronaviruses, which involves a specific virus receptor molecule on the cell surface. There is a rough correlation between receptor expression and the susceptibility of a cell type to virus infection (Benbacer *et al.*, 1997).

The receptors for TGEV, FIPV and HCoV have been identified as aminopeptidase N (APN) of the porcine, feline and human species, respectively (Delmas *et al*., 1992; Yeager *et al*., 1992; Benbacer *et al.*, 1997). APN is a member of the membrane-bound metallopeptidase family and is widely distributed on diverse cell types. It is particularly expressed on the brush border membrane of enterocytes.

TGEV also binds to a recently described 200kDa protein at the surface of the enterocytes on the villi of the small intestine (Weingartl & Derbyshire, 1994). The virus is resistant to the low pH of the stomach and passes to the small intestine where it infects the columnar epithelial cells covering the distal portion of the villi in jejunum and ileum. Coronavirus receptors are discussed in tabular form in Table 1.1.

1.1.4 Effect of age on infection

There is a clear relationship between TGEV pathogenicity and the age of the infected animal. The transmission of TGEV in older animals is by ingestion of contaminated material and in young piglets by the lactating sows. Contamination originates from faeces, milk and aerosols

generated in the respiratory tract. The adult swine needs 10^4-fold more virus infection than that required infecting a neonate (Regula *et al*., 2000).

Table* 1.1. Coronavirus receptors

Virus	Receptor	Distribution of receptor	Tropism of the virus	Associated diseases
(*Serogroup 1*)				
HCoV 229E	APN	Widely	Respiratory and	Common cold, respiratory,
TGEV (pig)	APN	Widely	enteric tract	enteric, hepatitis, neurologic,
FIPV (cat)	APN	Widely	hepatocytes	enteric infection
CCV (dog)	APN		Respiratory tract	Common cold
(*Serogroup 2*)				
HCoV OC43	Sialic acid	Widely	Respiratory tract	Common cold, respiratory,
MHV (mouse)	Bgp (Ig superfamily)	Widely		enteric, hepatitis, neurologic
BCoV (cow)	9-O-acetyl-neuraminic acid	Widely		enteric infection
(*Serogroup 3*)				
IBV (chicken)	?			Respiratory, hepatitis

*Table taken from the review article by Schneider-Schaulies (2000).

1.1.5 Penetration and uncoating

The mechanism of coronavirus entry into target cells is still under debate. Early electron microscopic studies showed virus particles inside lysosome-like vesicles near plasma membranes. This suggested that the virus enters the cell by endocytosis ('viropexis') (David-Ferreira & Manaker, 1965). Other studies suggested that the virus enters cells by direct fusion between virion and the plasma membranes (Dougheri *et al*., 1976). The exact mechanism of virus entry may depend on cell types and virus strain. The mechanism of virus uncoating, *i.e*., the release of virion RNA from the nucleocapsid, is also unclear.

1.2 Genome structure of coronaviruses

The coronavirus genome (Fig. 1.1) is a non-segmented, single-stranded, positive-sense RNA that is polyadenylated, and capped. Its size is about 27 to 32 kilobases (Lai & Cavanagh, 1997), which is remarkably large, compared to the other RNA viruses. The coronavirus

replicase gene alone (20-22 kb) is about the same size as an entire picornavirus (~8 kb) and vesicular stomatitis virus (~11 kb) genomes added together. In fact, coronaviruses have the largest viral RNA genome known to date. It has been speculated that the large size of the viral RNA genome employs special mechanisms of RNA synthesis to counter the deleterious effects of possible errors during RNA synthesis. The genomic RNA functions as mRNA and is infectious. It contains about 7-8 functional genes, 4 or 5 of which encode structural proteins. The structural proteins that are found in all coronaviruses, are encoded by genes located in the genomic RNA in the order 5'-polymerase-(HE)-S-E-M-N-3', with a variable number of other, mostly non-structural and largely non-essential, genes interspersed among them. The architecture of these non-structural protein genes varies significantly among different coronavirus species.

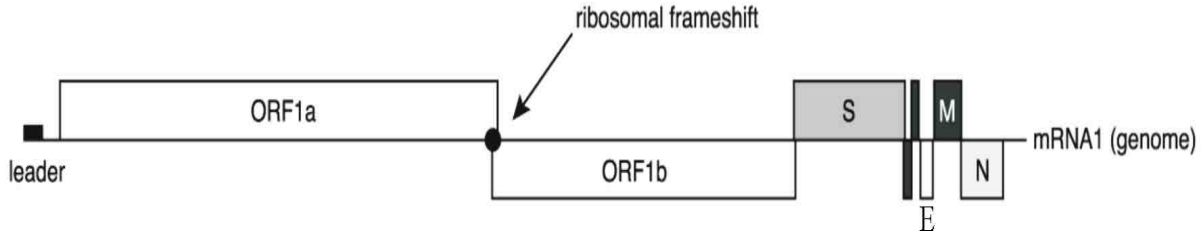


Fig. 1.1 Overview of the coronavirus genome organization. The 5' end of the genome represents the common leader sequence that is also present on each of the subgenomic mRNAs. The spike protein (S), integral membrane protein (M), envelope protein (E), and nucleocapsid protein (N) are also shown. The replicase gene is comprised of ORFs 1a and 1b, where the downstream ORF 1b is expressed by (-1) ribosomal frameshifting.

The replicase gene, which encodes all the functions required for viral replication and transcription (Thiel *et al*., 2001), occupies the 5'-terminal two-thirds of the genome and is comprised of two open reading frames (ORFs). In relation to the ORFs that encode non-structural proteins, the coronavirus genome is dominated by the replicase gene, which consists of two large ORFs (1a, 1b) encompassing approximately 20,000 nucleotides toward the 5' end of the genome. Expression of the downstream ORF is mediated by (-1) ribosomal frameshifting. The replicase gene thus encodes two large polyproteins, pp1a (450kDa) and pp1ab (750kDa) that are co- and post-translationally processed by viral proteinases to

produce the functional subunits of the viral replication complex (for reviews, see de Vries *et al*., 1997; Ziebuhr *et al*., 2000). Studies in the coronavirus system consistently revealed that the coronavirus main proteinase, M^{pro} (also called 3C-like proteinase, $3CL^{pro}$), cleaves pp1a and pp1ab at 11 conserved sites (Ziebuhr *et al*., 2000).

The divergence from a common ancestor and also RNA recombination appears to be a major driving force in the evolution of coronavirus genomes. This appears to have introduced plasticity to the coronavirus genome that is exceptional, even among RNA viruses. Probably because of the large size, coronaviruses have evolved many genetic mechanisms (e.g. genetic recombination and generation of defective interfering (DI) RNAs), to maintain their genetic stability.

1.2.1 Major gene products

Most groups of positive-stranded RNA viruses of animals produce either a single polyprotein or separate non-structural and structural precursor polypeptides that are subsequently cleaved by virus-encoded or host-encoded proteinases to produce functional subunits (Dougherty & Semler, 1993). In contrast, the nidovirus structural proteins that are encoded in the 3'-proximal region of the genome are individually expressed from a nested set of subgenomic mRNAs generated by a unique discontinuous transcription mechanism (Spaan *et al.*, 1983; Lai *et al.*, 1984; van Marle *et al.*, 1999).

The spike glycoprotein (S) is the heavily glycosylated outermost component of the corona virion, and has two biological activities important for the virus. It is responsible for the attachment of the virus to cells (Collins *et al.,* 1982; Godet *et al*., 1994; Kubo *et al.*, 1994) and for instigating the fusion of the virus envelope with cell membranes. The S protein is large, ranging from some 1160 (IBV) to 1452(FCV) amino acids. This protein soars some 20 nm above the virion envelope, giving the virus a 'solar corona-like' appearance under negatively stained electron micrographs (Fig. 1.2), from which the name of this virus family is derived.

The integral membrane glycoprotein (M) is one of the structural proteins that are essential for the production of coronavirus-like particles. The M polypeptide comprises 225-230 amino acids (MW ~ 25 kDa); some members of the TGEV group have additional 30 or so residues at the amino terminus, forming a cleavable membrane insertion signal. The protein has three membrane-spanning regions in the amino-terminal half. It probably plays a role in viral pathogenesis.

The hemagglutinin-esterase glycoprotein (HE) of approximately 424 amino acid residues (65 kDa) has been detected in virions of HEV, MHV, HCoV OC43, BCoV, and TCV. Those coronaviruses that contain HE in their virions cause hemagglutination much more efficiently than those that do not. As its name implies, the HE protein also has esterase activity, specifically, it is a neuraminate-O-acetylesterase. The functional significance of HE for coronaviruses is not known, and only BCoV requires HE for infectivity.

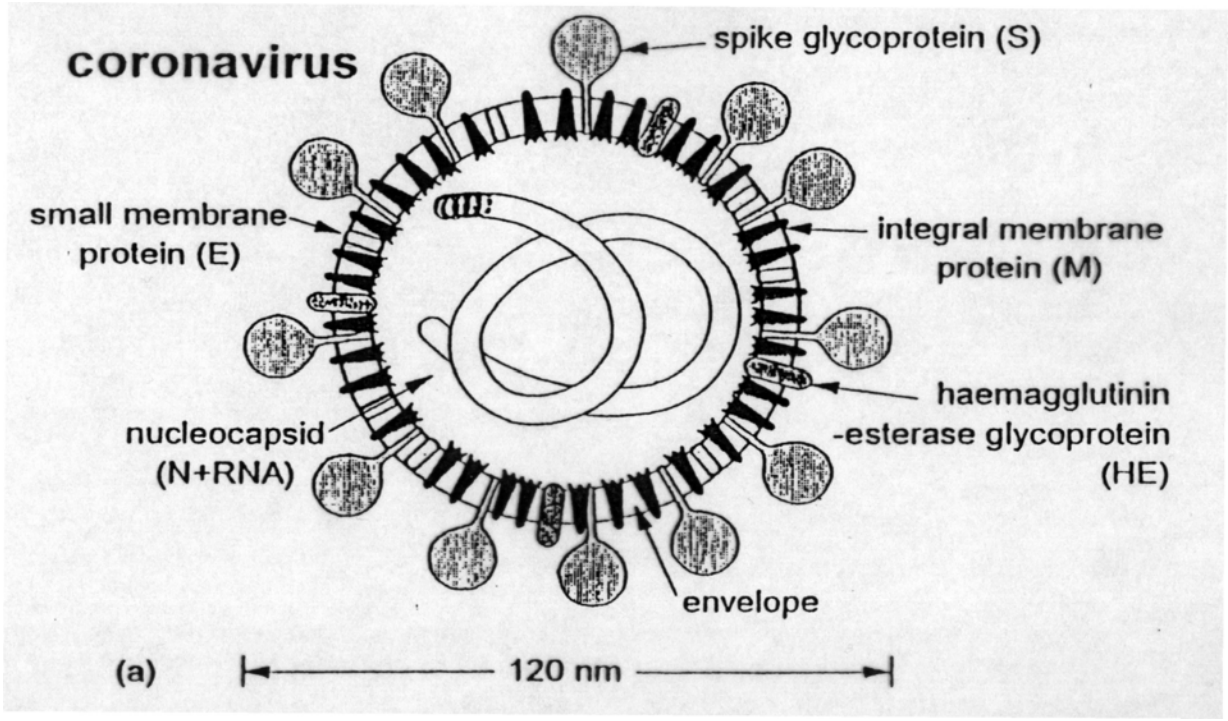


Fig. 1.2 Coronaviruses are pleomorphic but roughly spherical enveloped particles. They are about 120-160 nm in diameter with a characteristic 'fringe' of surface projections (20 nm long). Diagram taken from the text book ‘The Molecular Biology of Coronaviruses’ by Lai and Cavanagh.

The small membrane protein (E) is an additional virion protein. It plays an essential role in virion assembly. The E protein in the cell is localized in the perinuclear region, with some migrating to the cell surface (Godet *et al.*, 1992, Yu *et al*., 1994). The E proteins vary from 84 to 109 amino acid residues, corresponding to molecular weights of 9100 to 12,400 (Siddell, 1995).

The N protein is a 50 – 60 kDa phosphoprotein which, together with the genomic RNA, forms a helical nucleocapsid (RNP). The RNP of coronaviruses has been reported variously to be from 9-11 to 14-16 nm in diameter. The N protein varies from 377 to 455 amino acid residues in length. It is highly basic and has a high (7-11%) serine content, making it a potential target for phosphorylation.

Although from the same family, the coronavirus and torovirus gene products, with some limited exceptions, lack sequence similarity. The sizes of the nucleocapsid protein are as different as ~ 60 kDa and ~ 18 kDa, and the shape of the helical nucleocapsid structure is extended or tubular for the two viruses, respectively. The leader sequence is absent at the 5' end of torovirus mRNAs.

1.3 The RNA virus proteinases

The RNA virus proteinases have number of features: (i) they may function as discrete proteins or, more commonly as proteolytic domains of larger forms, (ii) these larger forms themselves may represent alternative processing products, (iii) proteinase activities may be different depending upon which type of proteinase domain is present, (iv) as complexes (by binding other virus proteins or RNA), their activity or substrate specificity may be modified, (v) virus-encoded proteinases may cleave host-cell proteins modifying or inhibiting host-cell functions, (vi) their activation may be delayed until a special environment is encountered (*i.e*. during capsid morphogenesis) and (vii) Autocatalytic cleavage can occur *in cis* or *in trans*, this means, they may cleave in an intramolecular manner (in *cis*), or in intermolecular manner (in *trans*) - although the catalytic mechanism of such cleavages is the same. Processing reactions occurring in *cis* are rapid as compared to *trans* reactions. The *cis* cleavage site may be proximal or distal, whereas *trans* reactions follow second-order kinetics

and are more sensitive to inhibitors and to sequence variations flanking the scissile amino-acid pair. Interestingly, they may exhibit highly specific and regulated proteolysis of other virus protein precursor molecules and may control the biogenesis of different, alternative, functions from the same precursor (e.g. picornavirus). Additionally, they may have specific cellular protein targets which, when cleaved, result in a modification of cellular macromolecular processes.

1.4 Accessory proteinases

All coronaviruses encode between one and three 'accessory' proteinases, which are distantly related. TGEV and HCoV encode two accessory proteinases. They are called coronavirus papain-like proteinases $PL1^{pro}$ and $PL2^{pro}$. These accessory proteinases recognize one or two sites that are located in the N-terminal half of the replicase polyproteins, cleave peptides that have small residues at the scissile bond, have a catalytic dyad involving cysteine and histidine (Ziebuhr *et al*., 1997b) and may employ a variant of the α+β fold that is conserved in the papain class of proteinases.

In coronaviruses accessory proteinases, the catalytic Cys and His residues are separated by amino-acid segments almost twice the size as compared to other papain-like proteinases, making these proteinases the largest in the class of RNA virus proteins. They have eight conserved residues, out of which the catalytic Cys-His dyad as well as three cysteines are involved in the formation of a zinc finger. The conservation of this structural element embedded in the central region of these enzymes clearly discriminates the accessory proteinases of coronaviruses from their arterivirus counterparts.

Only very limited information is currently available on the requirements of the accessory proteinase-mediated cleavages in other coronaviruses. The substrate specificity of the HCoV $PL1^{pro}$ was determined by sequence analysis of the carboxyl-terminal cleavage site of p9 (Herold *et al.*, 1998). Gly was found at the P1 position and the P1' and P5 positions were occupied by small uncharged (Asn-112) and basic (Lys-107) residues, respectively. In common with the coronavirus main proteinases, the coronavirus accessory proteinases have also been shown to deviate significantly from the prototypic cellular enzymes. The accessory

nidovirus proteinases mediate only very few cleavages in the relatively divergent, amino-terminal portion of the replicase polyproteins. In contrast, the main proteinases are responsible for the extensive proteolytic processing of the so-called 'core replicase' (Snijder & Spaan, 1995).

1.5 Structural aspects

Sequence comparisons have revealed that the two hydrophobic domains, HD1 and HD2, are situated on either side of coronavirus M^{pro} (Gorbalenya *et al.*, 1989*a*; Lee *et al.*, 1991; Herold *et al.*, 1993; Eleouet *et al.*, 1995); these domains are also conserved in arteriviruses. The data from *in vitro*-translation experiments showed that microsomal membranes are required for the efficient autoproteolytic processing of the M^{pro} from HD1 and HD2, most likely by assisting in the proper folding of these proteins (Tibbles *et al.*, 1996; Pinon *et al.*, 1997; Schiller *et al.*, 1998). On the other hand, after being released from the polyprotein, the M^{pro} activity does not depend on membranes, or any other cofactor(s), at least for its proteolytic activity *in vitro* (Ziebuhr *et al.*, 1995). It has been suggested that HD1 and HD2 may contribute to the intracellular localization of the M^{pro} itself and, possibly, of the virus replication complex in general (Gorbalenya *et al.*, 1989a). Coronavirus replication takes place at intracellular membranes and a large number of non-structural, replicase gene-encoded proteins contribute to the formation and function of the coronavirus replication complex.

The coronavirus M^{pro}s are the largest proteinases of known RNA viruses. They consist of 302–307 amino acid residues, whereas the prototypic poliovirus 3C proteinase contains only 182 residues. The unique, C-terminal domain of approximately 100 amino-acid residues that appears to be required for proteolytic activity is responsible for the size difference. Therefore, a large number of different carboxyl-terminally truncated versions of the HCoV M^{pro} are inactive in assays using synthetic peptides (Ziebuhr *et al.*, 1997a). Also, the removal of 28 carboxyl-terminal amino acids from the MHV M^{pro} abolishes its activity in an *in vitro* translation system (Lu & Denison, 1997). In apparent contrast to the HCoV and MHV data, it was recently shown that a recombinant form of the IBV M^{pro} tolerated the introduction of six consecutive His residues near its carboxyl terminus without loss of activity (Tibbles *et al.*, 1999). In this experiment, the His tag (His_6) was placed into the

nearest region of predicted hydrophobicity to the C-terminal processing site to minimize interference with the sequences involved in processing. The histidines were kept exposed to interact with a purification matrix. There are several speculations on the function of the carboxyl-terminal region (Zeibuhr *et al.*, 2000), *e.g.* (i) maintenance of the overall fold of the enzyme, (ii) involvement in catalysis or (iii) substrate recognition, and (iv) a non-proteolytic function.

1.6 TGEV and HCoV M^{pro}s

The viral proteins required for TGEV and HCoV genome replication and transcription are encoded by the replicase gene (Eleouet *et al.*, 1995; Penzes *et al.*, 2001). This gene encodes two replicative polyproteins, pp1a (447 kDa) and pp1ab (754 kDa), that are processed by virus-encoded proteinases to produce the functional subunits of the replication complex (reviewed in Ziebuhr *et al.*, 2000). The name "3C-like proteinase" was originally introduced because of similar substrate specificities of the coronavirus M^{pro} and picornavirus 3C proteinases ($3C^{pro}$) and the identification of Cys as the principal catalytic residue in the context of a predicted two-β-barrel fold (Gorbalenya *et al.*, 1989a,b). Meanwhile however, several studies have revealed significant differences in both the active sites and domain structures between the coronavirus and picornavirus enzymes (Liu & Brown, 1995; Lu & Denison, 1997; Ziebuhr *et al.*, 1997a, 2000; Hegyi *et al.*, 2002). Also, the crystal structures reported for a number of picornavirus 3C proteinases (Allaire *et al.*, 1994; Matthews *et al.*, 1994; Bergmann *et al.*, 1997; Mosimann *et al.*, 1997;) have not been useful in predicting the three-dimensional structures of coronavirus main proteinases. Because of the large phylogenetic distance between the two groups of enzymes, we are using the term coronavirus M^{pro} rather than $3CL^{pro}$.

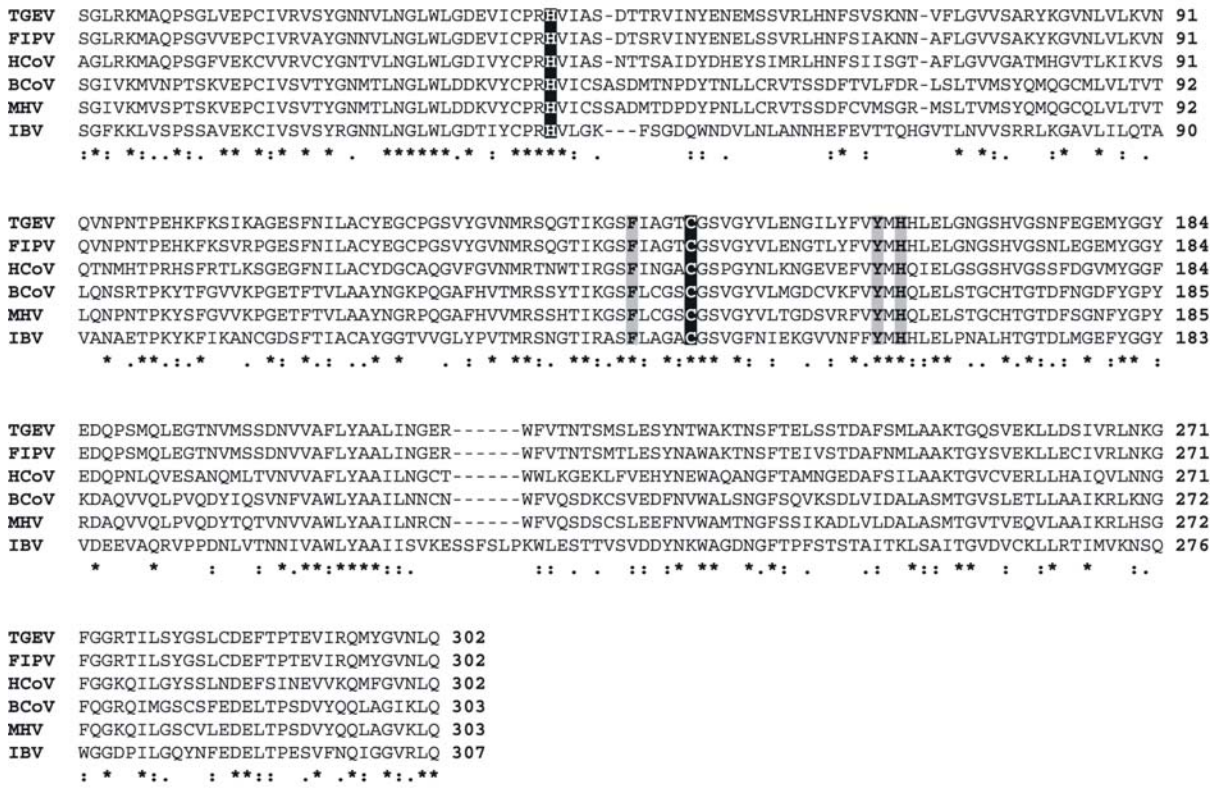


Fig. 1.3 Sequence comparison of coronavirus main proteinases. The alignment was produced using CLUSTAL X, version 1.81 (Thompson *et al.*, 1997). The sequences of FIPV (strain 79-1146), HCoV (strain 229E), BcoV (isolate LUN), MHV (strain JHM) and IBV (strain Beaudette) were derived from the replicative polyproteins of the respective viruses whose sequences are deposited in the GenBank database (FIPV, AF326575; HCoV, X69721; BCoV, AF391542; MHV, M55148; IBV, M95169; TGEV, AJ271965).

Sequence comparisons (Fig. 1.3) and experimental data obtained for other coronavirus homologues allows prediction that the mature form of the TGEV M^{pro} is released from pp1a and pp1ab by autoproteolytic cleavage at flanking Gln↓(Ser,Ala) sites (Eleouet *et al.*, 1995; Hegyi & Ziebuhr, 2002). Accordingly, these M^{pro}s have 302 amino-acid residues that correspond to the pp1a/pp1ab residues 2879 to 3180 for TGEV and residues 2966 to 3267 for HCoV. *In vivo* and *in vitro* analyses of IBV, MHV, and HCoV M^{pro} activities have shown consistently that these proteinases cleave the replicase polyproteins at 11 conserved sites and,

therefore, it seems reasonable to conclude that the M^{pro}-mediated processing pathways are conserved in all coronaviruses.

Previous theoretical studies and experimental data have led to the following conclusions (Bazan & Fletterick, 1988; Gorbalenya *et al.*, 1989a,b; Liu & Brown, 1995; Lu *et al.*, 1995; Lu & Denison, 1997; Ziebuhr *et al.*, 1995, 1997a, 2000; Seybert *et al.*, 1997; Ziebuhr & Siddell, 1999; Ng & Liu, 2000; Hegyi *et al.*, 2002):

i) Coronavirus main proteinases employ conserved Cys and His residues in the catalytic site. In TGEV and HCoV M^{pros}, these are Cys144 and His41. There has been some debate on the existence of a third residue in the catalytic centre. Gorbalenya *et al*. (1989b) predicted a catalytic triad consisting of His 2820, Cys2922 and Glu2843, for coronavirus IBV M^{pro}. In common with picornavirus 3C proteinases, the catalytic center of the coronavirus M^{pro} is predicted to be embedded in a chymotrypsin-like, two-β-barrel structure in which cysteine (rather than serine) serves as the principal nucleophile.

ii) These proteinases have well-defined substrate specificities. All known cleavage sites contain bulky hydrophobic residues (mainly Leu) at the P2 position, Gln at the P1 position, and small aliphatic residues at the P1' position.

iii) They possess a large C-terminal domain of about 100 amino-acid residues that is not found in other RNA virus 3C-like proteinases. The characterization of recombinant proteins, in which 33, 28, and 34 C-terminal amino acid residues were deleted from the IBV, MHV and HCoV main proteinases, respectively, resulted consistently in dramatic losses of proteolytic activity, suggesting that the C-terminal domain of M^{pro} contributes to proteolytic activity through undefined mechanisms.

1.6.1 Catalytic center and substrate specificity

The sequence similarities of the M^{pro}s to prototypic picornavirus 3C proteinases are limited to the catalytic region in the profile-versus profile dot-plot cross analysis (Ziebuhr, 2000). The catalytic residues of 3C and 3C-like proteinases are superimposed upon a two-β-barrel structure consisting of 12 antiparallel β-strands (Bazan & Fletterick, 1988; Gorbalenya *et al.*, 1989*a*; Allaire *et al.*, 1994; Matthews *et al.*, 1994; Mosimann *et al.*, 1997). In 3C and 3C-like proteinases, Cys replaces the nucleophilic Ser and, in a subset of viruses, Glu replaces the

Asp of the catalytic triad found in cellular proteinases (Bazan & Fletterick, 1988; Gorbalenya *et al.*, 1989*a*; Matthews *et al.*, 1994). The coronavirus M^{pro} seems to lack a conserved acidic residue that would be equivalent to the catalytic Asp (Glu) of 3C proteinases. The available data strongly suggest that the conserved His and Cys residues represent the general base and nucleophile, respectively.

The coronavirus M^{pro} displays additional features that clearly separate it from other virus-encoded 3C-like proteinases, including the arterivirus main proteinase. For example, it employs a novel version of the substrate-binding pocket 'core' motif, which is characteristically Gly-X-His for most other 3C/3C-like proteinases. Thus, the Gly residue of this motif (Bazan & Fletterick, 1988; Gorbalenya *et al.*, 1989*a*, *b*) is conserved in the vast majority of serine and cysteine proteinases with chymotrypsin-like (CHL) folds and only very few proteinases tolerate substitutions with small amino acids (Ala or Cys) at this position. This conservation pattern indicates a strong selection pressure with regard to the space that this specific residue occupies. In contrast, in all coronavirus $3CL^{pro}$ domains studied so far, Gly appears to be replaced by Tyr (Gorbalenya *et al.*, 1989*b*; Lee *et al.*, 1991; Herold *et al.*, 1993; Eleouet *et al.*, 1995) (Fig. 1.3). The Tyr residue of the Tyr-X-His motif has not yet been probed by mutagenesis. However, the replacement of the His residue (His-3127) by Ser completely abolished the proteolytic activity of the HCoV M^{pro} (Ziebuhr *et al.*, 1997b). This inactivation was selective since a similar replacement of His-3136, another conserved His residue in this region, was not so detrimental (Ziebuhr *et al.*, 1997a). Thus, the importance of the Tyr-X-His motif has been confirmed, implying that coronaviruses may indeed have accepted a Gly-to-Tyr replacement during evolution.

The above data are also compatible with a model, originally developed and substantiated for other 3C/3C-like proteinases (Bazan & Fletterick, 1988; Gorbalenya *et al.*, 1989*a*; Allaire *et al.*, 1994; Matthews *et al.*, 1994; Mosimann *et al.*, 1997), that implicates His-3127 (and its counterparts in other coronaviruses) in the formation of hydrogen bonds to the P1 glutamine side chain of M^{pro} substrates (Gorbalenya *et al.*, 1989*b*).

The substrate specificity of M^{pro} resembles that of many other 3C/3C-like proteinases (Kräusslich & Wimmer, 1988; Dougherty & Semler, 1993 ; Blom *et al.*, 1996) in so far as the P1 position of the substrate is exclusively occupied by Gln and small, aliphatic residues (Ser, Ala, Asn, Gly and Cys) are found at the P1' position. However, Asn and Cys are most uncommon as P1' residues outside of the coronavirus family, although a P1' Asn is found in rhinoviruses (Blom *et al.*, 1996) and, in a mutagenesis study, Cys proved to be a tolerable substitution in one of the encephalomyocarditis virus (EMCV) $3C^{pro}$ sites (Parks *et al.*, 1989). In four different coronaviruses, one M^{pro} cleavage site consistently contains Asn at the P1' position (Liu *et al.*, 1997; Lu *et al.*, 1998; Ziebuhr & Siddell, 1999) and, for MHV, a P1' Cys residue was predicted for another site (Lee *et al.*, 1991).

Upon comparison of a large number of cleavage sites, most of which have experimentally been confirmed for at least one coronavirus, it is evident that in addition to P1 and P1', the P2, P3, P4, P2' and P3' positions have a restricted variability. Among these, the P2 and P4 positions are most conserved with bulky hydrophobic residues (mainly Leu) at P2 and Val, Thr, Ser (and Pro) at P4 being clearly favoured (Ziebuhr *et al.*, 2000). The efficiency of cleavage at specific sites is likely to be determined by the exact composition of the sites, since synthetic peptides mimicking different cleavage sites were processed in competition experiments at significantly different rates by the HCoV M^{pro} (Ziebuhr & Siddell, 1999). In view of these data, it seems likely that together with the accessibility of potential cleavage sites in the context of the polyprotein, the properties of the cleavage sites themselves contribute significantly to the coordinated, temporal release of specific polypeptides from the replicase polyproteins. This might lead to the (irreversible) activation or inactivation of specific functions in the course of the virus life cycle, as has been demonstrated for a number of other positive-stranded RNA viruses.

1.7 Aims and objectives of this work

Like many other viral proteinases, the TGEV and HCoV proteinases, are highly effective regulators of virus replication and, indirectly, possibly even of virion biogenesis (van Dinten *et al*., 1999). Herein, a crystallographic approach is undertaken with wild-type TGEV M^{pro}, wild type and several mutants of HCoV M^{pro} and inhibitor complexes in order to arrive at a

better understanding of the enzyme's functional properties in protein biosynthesis. A prerequisite for this, is the availability of good crystals, which is often a major impediment to this type of work. Therefore, the first aim of this work was to determine the optimal conditions for crystallization of TGEV and HCoV proteinases. Once the crystallization conditions were optimized and diffractable crystals were obtained, the next task was to solve the phase problem, to allow construction of three-dimensional models. Many biochemical observations have been made over the past years in relation to the enzymes behavior, and require rationalization on the basis of its structure. The active-site residue Cys144 requires precise investigation vis-à-vis catalytic and substrate-binding/cleavage mechanisms employed by the main proteinases of the coronavirus. The TGEV and HCoV M^{pro}s are peculiar for their unique C-terminal domain, the function of which is unknown. The novel Tyr-X-His motif remains to be structurally explained; there are numerous suggestions as to the RNA binding role of these proteins, and a plethora of other unanswered questions. Since this work represents the first effort towards the three-dimensional structure of any protein of the related families *Coronaviridae* and *Arteriviridae*, this work will enable numerous other aspects of coronaviruses to be studied, including structural details, the role of many conserved residues in maintaining active site geometry, the catalytic dyad, the catalytic mechanism, details of substrate specificity, and ultimately the design of inhibitors and targets for therapeutic intervention.

2. *MATERIALS AND METHODS*

2.1 Materials

2.1.1 Equipments

A list of laboratory instruments and devices used for this work is given below in Table 2.1.

Table 2.1 ***General laboratory devices and their manufacturers***

Equipment	**Manufacturer**
Protein purification/concentration/analysis	
Centrifuge – Heraeus Labofuge 400R	Heraeus Instruments (Hanau)
Centrifuge – Heraeus Biofuge plus	Heraeus Instruments (Hanau)
Spectrophotometer – UV Vis Spekol	Zeiss (Jena)
Analytical balance – Sartorius BP 210 D	Sartorius (Göttingen)
Table balance – Sartorius portable PT2100	Sartorius (Göttingen)
Water purification system – Milli-Qplus 185	Millipore (Eschborn)
pH Meter – CG 840 Schott	Schott (Mainz)
Gel electrophoresis system	Pharmacia (Freiburg)
Mass spectrophotometer (MALDI-TOF) – Biflex II	Bruker (Karlsruhe)
DLS instrument: DynaPro-801	Protein Solutions Ltd. (Buckinghamshire, UK)
Crystallization	
Incubator – EHRET KBK 4200	EHRET (Emmendingen)
Incubator – EHRET KBK 4600	EHRET (Emmendingen)
Microscope – Olympus SZH10 binocular	Olympus (Hamburg)
Microscope – Zeiss Stemi 1000 binocular	Zeiss (Jena)
Data collection	
X-ray generator, rotating anode – Nonius FR591	Nonius (Delft, The Netherlands)
Image plate detector – Mar 300	Mar Research (Hamburg)
Image plate detector – Mar 345	Mar Research (Hamburg)
Image plate detector – Dip 2030K	Nonius (Delft, The Netherlands)
Cryostat- Oxford controller 600 series	Oxford Cryosystems (Oxford, UK)
Air stream cooler – FTS TC84	FTS systems (Stone Ridge, USA)

X-ray mirror system – MAC-XOS	MacScience (Yokohama, Japan)
Synchrotron – EMBL Hamburg beamlines X11, BW7A	DESY (Hamburg)
Synchrotron – Elettra Light Source beamline 5.2 R	ELETTRA (Trieste, Italy)
Goniometer head – Charles Supper Standard	Charles Supper (Troy, USA)
Microscope – Leica MZ 8 binocular	Leica (Bensheim)
Computing	
Indy workstation	SGI (Mountain View, USA)
Onyx graphics workstation	SGI (Mountain View, USA)
O2 graphics workstation	SGI (Mountain View, USA)
Indigo2 graphics workstation	SGI (Mountain View, USA)
Origin 200 server	SGI (Mountain View, USA)
Digital 433au	Compaq (Houston, USA)
Data cartridges	Sony, Maxell (Japan)

2.1.2 Chemicals

A list of chemicals used for this work is given below in Table 2.2.

Table 2.2 Chemical items and their manufacturers

Chemical item	**Manufacturer***	**Chemical item**	**Manufacturer***
Ammonium sulfate	Merck	Izit™	Hampton Research
BSA	Aldrich	MPD	Merck
Carbonic anhydrase	Aldrich	NaCl	Fluka Chemie AG
Crystallization screen	Hampton Research	Paraffin oil	Merck
Deionized water	Millipore	PEG 6K, 10K	Fluka Chemie AG
Dioxane	Merck	SDS	Pharmacia Biotech
DTT	Merck	Silicon fluid 200/1 cS oil	Merck
Ethanol	Merck	Sinapinic acid	Aldrich
1,6-hexanediol	Merck	TLCK	Sigma
HEPES	Fluka Chemie AG	Tris	Merck

* more details are given in the text

2.1.3 Crystallization materials and cryo-tools

Centricon Plus-20 centrifugal filter devices and Ultrafree-MC filter units: Millipore (Bedford, USA), dialysis buttons, dialysis rings and slide-A-Lyzer dialysis cassettes: Pierce (Rockford, USA). Dialysis membranes and sample tubes: Roth (Karlsruhe, Germany), glass sample capillaries: GLAS (Berlin, Germany), highly liquid paraffin oil: Merck (Darmstadt, Germany), high vacuum grease: Dow Corning (Midland, USA). Magnetic base crystal caps, mounted cryoloops, 24-well linbro plates and VDX plates, 22 mm circular siliconized coverslips, sealing wax, Crystal Screen 1 and 2, crystal storage vials, cryo canes, magnetic crystal wands, curved vial clamps and micro tools: Hampton Research (Laguna Niguel, USA). Other basic chemicals were obtained from Sigma-Aldrich Chemie GmbH (Munich, Germany) and Fluka Chemie GmbH (Buchs, Switzerland).

2.1.4 Buffers and solutions

All buffers and solutions were prepared using deionised water at 20°C (Millipore water purification system).

Table 2.3 Buffer, protein storage solutions and oil

Buffer	HEPES pH 8.5 and pH 8.8
Protein storage solution (HCoV M^{pro})	11mM Tris.HCl pH 8.0, 200mM NaCl, 0.1mM EDTA, 1mM DTT
Protein storage solution (TGEV M^{pro})	11mM Tris.HCl pH 7.4, 120mM NaCl, 0.1mM EDTA, 1mM DTT
Oil mixture	50% paraffin and 50% silicon fluid 200/1 cS 60% paraffin and 40% silicon fluid 200/1 cS 70% paraffin and 30% silicon fluid 200/1 cS

2.1.5 Protein samples

Purified TGEV M^{pro}, HCoV M^{pro} wild type and HCoV M^{pro} Δ301-302 (deletion mutant) were kindly provided by Dr. J. Ziebuhr (Institute of Virology and Immunology, University of Würzburg, Germany). Low molecular-weight protein standards, Phastgel homogenous gel beds and 12.5% SDS buffer strips were from Pharmacia (Freiburg).

2.2 Methods

2.2.1 Proteins

The purification procedure for recombinant TGEV M^{pro} (residues 1 to 302) and HCoV $M^{pro}\Delta$301-302 (residues 1 to 300) is described by Ziebuhr *et al* (1997) and Hegyi *et al* (2002). The proteins were expressed in *Escherichia coli* as an MBP fusion protein and first purified by amylose affinity chromatography. The recombinant M^{pro} was then released from the fusion protein by factor Xa cleavage and purified to apparent homogeneity by hydrophobic interaction, ion exchange and gel filtration chromatography. The recombinant proteinases were finally concentrated to 12.5 and 15 mg/ml (Centricon-YM3, Millipore, Eschborn, Germany), respectively.

2.2.2 Selenomethionine-derivatized proteins

Practical impediments to structure solution from the native protein, compounded by its low sequence homology to other known proteinases, persuaded us to produce selenomethionine SeMet-substituted M^{pro} for phase determination using the MAD approach. The SeMet-derivatized TGEV and HCoV M^{pro} were also provided by John Ziebuhr. The plasmids, pET-TGEV M^{pro} and pET-HCoV $M^{pro}\Delta$301-302, were used to transform the methionine auxotrophic 834(DE3) *E. coli* strain (Novagen, Germany). The SeMet-substituted coronavirus proteinases were concentrated to 9.5 mg/ml and 7.1 mg/ml, respectively.

2.2.3 Determination of purity

Protein purity is a critical factor in crystallization experiments: proteins used for crystallization should be as pure as possible and completely homogeneous (McPherson, 1998). The purity of TGEV M^{pro} and HCoV M^{pro} was visually evident from 12.5% SDS polyacrylamide gels stained with Coomassie Blue (Lämmli, 1970) (Fig. 2.1).

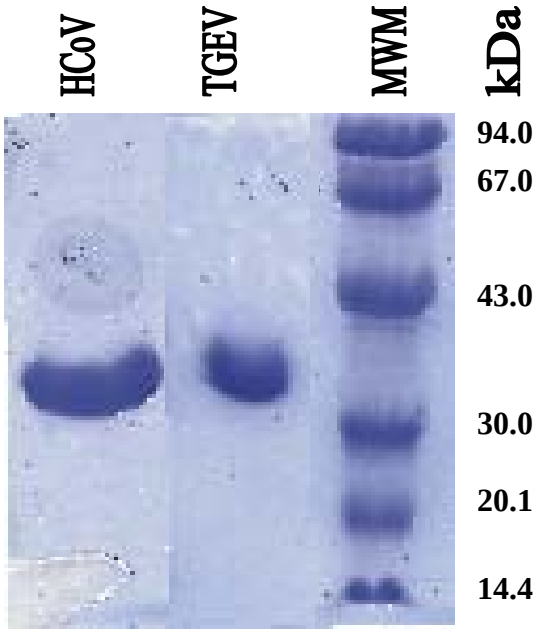


Fig. 2.1 SDS-PAGE gel analysis of the TGEV and HCoV M^{pro}. Both lanes show individual single bands of approximate 33 kDa molecular weight.

2.2.4 Characterization of purified TGEV M^{pro} and HCoV M^{pro}

2.2.4.1 MALDI Mass Spectroscopy

Molecular masses of the proteins were verified by MALDI (matrix-assisted laser desorption ionization) mass spectroscopy. Mass spectroscopy was performed by Dr. K. H. Gührs (IMB, Jena) and Dr. N. Oldham (MPI for Chemical Ecology, Jena) on protein under low salt concentration (native protein).

***TGEV M^{pro}*:** The protein samples were diluted to 5 pmol/µl with 0.1% trifluoroacetic acid (TFA), mixed with the matrix (saturated solution of sinapinic acid or α-cyano-4-hydroxycinnamic acid (Fluka, Germany), 66% acetonitrile in water containing 0.1% TFA), and analyzed with a Biflex II workstation (Bruker, Germany). Bovine serum albumin (MW = 66.431 kDa) was used as the internal mass standard. Protein solution and matrix were mixed in equal volume and 1 µl of the sample was applied to the target. Mass spectra were analyzed using the analysis software XMASS v5.0 (Bruker, Germany). The mass spectra were essential to check the selenium incorporation in SeMet-derivatized TGEV M^{pro} protein before setting up crystallization trials.

***HCoV M^{pro}*:** As the spectrum intensity of desired mass was quite low with the α-cyano-4-hydroxycinnamic acid, spectroscopy was done with at least two more matrices (sinapinic acid and 2,5-dihydroxybenzoic acid) and the internal mass standard carbonic anhydrase (MW = 30 kDa) was taken for both native as well as SeMet-derivatized Δ301-302 mutant of HCoV M^{pro}.

2.2.4.2 Dynamic Light Scattering

The oligomerization state of the protein in solution was analyzed by dynamic light scattering (DLS). In the DLS measurement, a beam of monochromatic light was directed through the sample to monitor fluctuations in the light intensity scattered by the protein molecules. From analysis of the data, the translational diffusion coefficient (D_T) of the protein particles in solution was obtained. Assuming Brownian motion, this coefficient was converted to the hydrodynamic radius, R_H, of the protein particles using the Stokes-Einstein equation ($R_H = k_bT/6\pi\eta DT$), where k_b represents Boltzmann's constant, T is the absolute temperature in Kelvin and η is the solvent viscosity. The hydrodynamic molecular weight of the protein particle can be estimated from the measured value of D_T using a calibration curve obtained from proteins of known mass. DLS analysis was performed at 20 ^oC using a DynaPro-801 (Protein Solutions Ltd., Buckinghamshire, UK). The protein samples at 2-5 mg/ml in 100mM Tris.HCl pH 7.4 (buffer A) were filtered and centrifuged according to the manufacturer's instructions in order to make the solution free from particulate matter. 15-20 readings per sample were taken. All DLS analysis was carried out using the Dynamics version 5.24.02 program (DynaPro control software, Protein Solutions Ltd., UK). A Spectra-Physics Millennia II laser operating at 532 nm was used as the light source. The data were analyzed using the DynaLS software as described by Moradian-Oldak *et al*. (1998). Default parameters were used in interpreting the statistical results.

2.2.5 Crystallization experiments

Crystallization of the protein under study is a prerequisite for the entire crystallographic work. After the purification of proteins, crystals may be obtained by carefully searching for suitable crystallization conditions, which include pH, buffer condition(s), temperature, precipitant(s), etc. as variables (McPherson, 1998). The most common set-up to grow protein

crystals is the hanging-drop technique, which was used in this work. The technique is based on vapor diffusion of water. A few micro-liters of protein solution is mixed with an about equal amount of reservoir solution containing the precipitants. A drop of this mixture is put on a siliconized microscope glass slide, which covers the depression in a tray. The depression is partly filled with the required precipitant solution (reservoir solution: approximately 1 ml). Applying grease to the circumference of the depression seals the chamber, before the glass slide is put into place (Fig. 2.2). As the precipitant concentration in the drop is lower than in the reservoir, water evaporates from the drop and diffuses into the reservoir. As a result, the concentration of both protein and precipitant in the drop slowly increases and crystals may form.

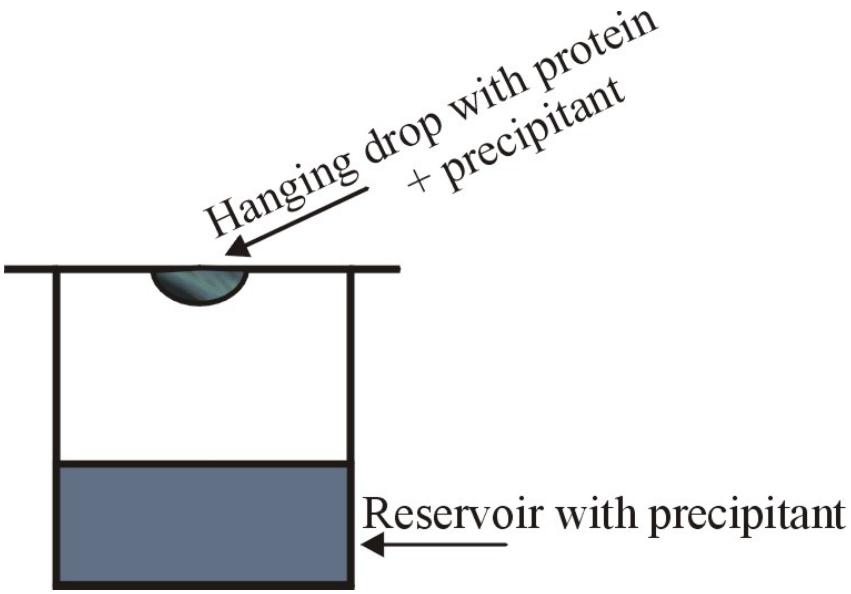


Fig. 2.2 Schematic diagram of a hanging drop setup.

***TGEV M^{pro}*:** Preliminary crystallization trials were conducted using Crystal Screens I and II and Additive Screens I and II (Hampton Research, California, USA). The hanging-drop vapor diffusion method was used at 4 ^{o}C to crystallize wild-type TGEV M^{pro}. A solution containing 12.5 mg/ml protein in 12mM Tris.HCl, 120mM NaCl, 0.1mM EDTA, 1mM DTT at pH 7.4 was used. The best crystals (Fig. 3.1A and B) were obtained by using a reservoir containing 100mM HEPES pH 8.8, 1.8M ammonium sulphate, 6% MPD, 5mM DTT, with 4% dioxane added after setting up the drop. Typically, a 2 µl aliquot of the protein solution was mixed with an equal volume of the reservoir solution and allowed to equilibrate against

1 ml of reservoir solution. Crystals of dimensions ~0.30x0.25x0.30 mm^3 grew in about 10 days. Crystallization plates set up using aged protein (of about 3 weeks after purification) gave better-quality crystals than those from freshly purified protein. This may be related to the oxidation of Cys144, where the difference density indicates the formation of the sulfinic acid ($–SO_2^-$) or sulfonic acid ($–SO_3^-$) derivatives in all TGEV M^{pro} monomers (Section 3.4.1).

Crystals of SeMet-derivatized TGEV M^{pro} were grown under the same conditions as for the native protein, but with 2M ammonium sulphate and 8% MPD. These crystals were extremely fragile and were difficult to be transferred to a storage buffer or cryogenic buffer without severe damage. Fortunately, they could be cooled to cryogenic temperatures after a quick rinse with mustard oil. They took two more days than the wild-type crystals to grow bigger and diffractable. The size of these crystals was approximately 0.25x0.25x0.20 mm^3.

***HCoV M^{pro}*:** Initial trials for the C-terminal deletion mutant (Δ301-302) of HCoV M^{pro} involved the hanging-drop vapor diffusion method at 4, 10 and 15 ^oC using Hampton Crystal Screens I and II. The protein concentration was 15 mg/ml in the protein storage solution (11mM Tris.Cl, 200mM NaCl, 0.1mM EDTA, 1mM DTT at pH 8.0). Each well of a Sarstedt tissue-culture plate contained 1 ml precipitant and above it 2 μl of a 1:1 mixture of protein with well solution.

A **B**

Fig 2.3 **A**. Native crystal of HCoV M^{pro} obtained by the hanging-drop method. The larger plates showed inferior diffraction quality. **B**. SeMet-derivatized crystals of HCoV M^{pro}.

The precipitant contained 15% PEG 6K, 4% 1,6-hexanediol, 5mM DTT, 2% dioxane and 100mM HEPES pH 8.5. Within three days, microcrystals appeared and in about one week, a few big plate-like crystals appeared at 10 ^{o}C temperature. The optimal size for X-ray diffraction of the crystal was ~0.3x0.1x0.05 mm^3. Larger crystals were more fragile and were leading to streaky reflections, whereas smaller ones were not diffracting better (Fig. 2.3A, B).

The SeMet-derivatized HCoV (Δ301-302) M^{pro} crystals were also obtained by using the hanging-drop method. The protein concentration was 7.1 mg/ml. The reservoir contained 100mM HEPES pH 8.5, 20% PEG (10K), 2% 1,6-hexanediol and 12% dioxane. The protein solution and reservoir were mixed in a 1:1 ratio for the crystallization drop and dioxane was added later to the reservoir solution, after setting up the drop. Crystals of about 0.2x0.2x0.05 mm^3 size appeared in about one week.

2.2.6 Characterization of protein crystals

There are a number of ways to distinguish crystals of salt from protein. (i) The light microscope can detect protein crystals in the small crystallization drop using birefringence. This way one can differentiate amorphous precipitate from microcrystals in a drop when viewed under a microscope. Precipitate does not have birefringent properties while all crystals except cubic ones do so. (ii) The second method to verify that crystals are proteinaceous, is to analyze them by SDS gel electrophoresis (Lämmli, 1970). (iii) The third method is the most commonly used one – Staining with Izit^{TM} ($C_{16}H_{18}ClN_3S$) or methylene green. This method was used to identify microcrystals of the TGEV and HCoV proteinases. If the crystal is a protein crystal, it turns blue, otherwise it stays opaque. This can simply be done by placing 1 µl of Izit^{TM} in the sample drop and waiting for 24 hours. Izit^{TM} is a small molecule dye that fills the solvent channels in protein crystals and binds with the protein molecules coloring the crystals blue. Salt crystals do not possess these large solvent channels thus Izit^{TM} cannot enter the crystal, leaving one with a clear crystal and a blue drop. Izit^{TM} works especially well for small microcrystals or for questionable precipitate. Another method is (iv) Poking the crystal with a very fine glass rod or glass pipette (drawn over a Bunsen burner) or a simple needle. This is the simplest and probably the best test. If the crystal is protein, it should disintegrate very easily. Protein crystals are more fragile and brittle than

salt crystals that are harder to break. (v) Setting up a "no protein" control drop with the buffer (+additives) the protein was in initially, to see if crystals grow without protein. Finally, the X-ray diffraction is the most reliable method to verify protein crystals.

2.2.7 Preparation of crystals for data collection

***TGEV M^{pro}*:** Elucidation of optimal cryo-conditions involved many trials with conventional cryoprotectants, none of which worked. TGEV M^{pro} crystals could not even survive the dry paraffin oil technique (Tunnicliffe and Hilgenfeld, 1999). The following hydrocarbon silicon oils were used in cooling experiments: Dow Corning 200/1cS fluid (BDH silicon products), highly liquid paraffin (Merck), viscous paraffin (Merck), vacuum-pump oil (Savant SPO1), rotary vacuum-pump oil (Leybold 17702), Al's oil (Hampton Research), mineral oil (Sigma), and different combinations of paraffin and silicon oils. Interestingly, a quick rinse at 4^{o}C in unrefined mustard oil (mustard seeds oil, Delhi, India) was more successful. Crystals were immediately frozen in liquid nitrogen.

***HCoV M^{pro}*:** These crystals did not need any external cryoprotectant and could be put straight from the drop into the stream of liquid nitrogen using a nylon loop (Hampton Research). The precipitant contained 20% PEG 10K, which acts as a cryoprotectant itself.

2.2.8 Diffraction data collection

***TGEV M^{pro}*:** Diffraction data were collected from crystals held in a stream of cooled nitrogen gas at 100 K (Oxford Cryosystems, Oxford, United Kingdom). The loop-mounted crystals were fragile and suffered from radiation damage when exposed for longer times to X-rays. Native data were initially collected to a Bragg spacing of 3.5 Å on an FR591 rotating copper anode X-ray generator (Nonius, Delft, The Netherlands) operated at 40 kV and 100 mA, using a 30-cm MarResearch image plate (X-ray Research, Hamburg, Germany). Data beyond 1.95 Å resolution were subsequently collected (Table 3.2) at 100 K on the X-ray diffraction beamline at ELETTRA, Trieste, using a Mar165 CCD detector (detector to crystal distance 140 mm).

***HCoV M^{pro}*:** Diffraction data were collected from a crystal held in a stream of nitrogen gas at 100 K (Oxford Cryosystems) using Cu-Kα.radiation generated at 40 kV and 100 mA by an FR591 rotating anode X-ray generator (Nonius, Delft, The Netherlands), equipped with a 30-cm image plate (Mar Research, Hamburg, Germany). The data extended to approximately 3.5 Å resolution, at an hour exposure per degree and crystal-to-detector distance of 250 mm. The low-resolution limit with CuKα radiation made the use of synchrotron radiation mandatory.

In order to optimize the use of synchrotron radiation, data for both TGEV and HCoV M^{pro}s were collected in a way (calculated by the program STRATEGY (Ravelli *et al*., 1997)) producing the complete data set with a minimal number of images. The space group and cell dimensions from the crystals were derived using the autoindexing routine in DENZO (Otwinowski & Minor, 1997). Data integration/reduction and scaling were performed using the programs DENZO and SCALEPACK (Otwinowski & Minor, 1997).

2.2.9 Initial attempts to solve the TGEV M^{pro} structure

Due to the low sequence similarity to other proteinases, the structure could not be solved using conventional molecular replacement techniques. Soaking and cocrystallization mainly with Hg, Pt, Au, Ir, Cd, Br compounds and xenon occlusion at 10 bars pressure were tried. Unfortunately, none of the 35 heavy atom compounds tried for TGEV M^{pro} did successfully bind to the protein crystal, including xenon and Br. Because it was a long list of trials, therefore, the prescreening method of Boggon and Shapiro (2000) was used to run a native gel before setting up the crystallization plate. 1 µl of protein with 1 µl of heavy atom containing solution was taken and incubated for about 10 min followed by running a native gel to look for band shifts. Few mercury compounds and iridium-soaked proteins showed a band shift (Fig. 3.4) in the native gel but after cocrystallization and soaking, it was found that only mercuric acetate bound partially (the crystals with this compound only, were quite isomorphorous to the native ones). However, it was impossible to solve the structure by the conventional isomorphous replacement method due to the non-isomorphism between different crystals. It was not possible to scale data sets from native and derivatized crystals together. Consequently, MAD data sets were collected to 2.9 Å resolution from crystals of SeMet-TGEV M^{pro} at a tunable synchrotron beamline.

2.2.10 Multiwavelength anomalous dispersion (MAD)

MAD Theory: The quantity usually measured in relation to each X-ray reflection is the intensity, which is proportional to $|F(HKL)|^2$ and hence it is |F(HKL)| that is determined experimentally. This quantity may be called the 'geometric structure factor' as it depends only on the positions of the atoms and not on any differences in their scattering behavior. If the nature of the scattering, including any phase change, is identical for all atoms, then |F(hkl)| = |F(-h-k-l)|. This is known as Friedel's law. However, when the energy range of the X-rays used is tuned to match that for the absorbance of a heavy atom, a breakdown of Friedel's law results. This phenomenon is known as anomalous scattering and is the basis of MAD experiments. In the presence of anomalous scattering atoms in the protein crystal, one can utilize two types of signal to calculate phases: the dispersive difference signal (between wavelengths, due to the contribution of Δf' to the structure factor), and the anomalous signal (between Bijvoet pairs, the intensity of Friedel-mate reflections is different due to the contribution of Δf"). These signals are used in a MAD experiment using tunable synchrotron radiation, to maximize both the dispersive and anomalous differences. This requires collection of at least three data sets, one at each wavelength. In the MAD method, the wavelength dependence of the anomalous scattering is used.

The principle of this method is rather old but it was the introduction of tunable synchrotron radiation sources that made it a technically feasible method for protein structure determination. Hendrickson and colleagues (1988) were the first to take advantage of this approach and use it for solving protein structures. Hendrickson showed that the presence of one selenium (Se) atom (atomic number 34) in a protein of not more than approximately 150 amino-acid residues is sufficient for a successful application of MAD; however, this depends very much on the quality of the data. With more Se atoms the size of the protein can, of course, be larger.

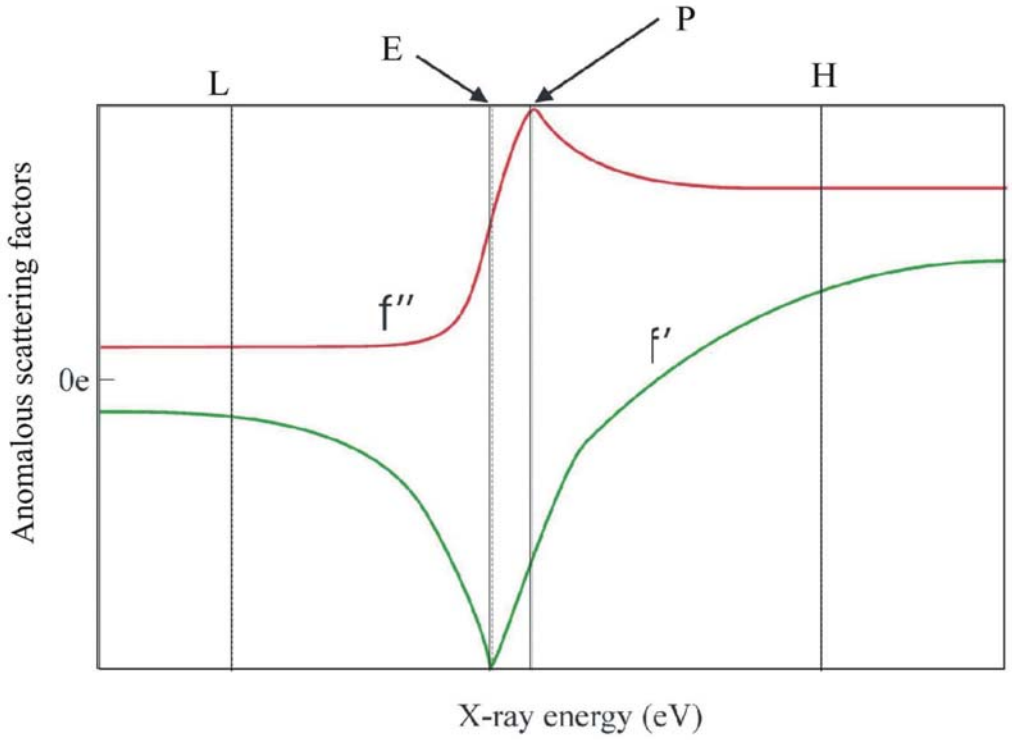


Fig. 2.4 Schematic of experimental values for Δf ' and Δf " as a function of X-ray energy.

The wavelengths have to be carefully chosen to optimize the difference in intensity between Bijvoet pairs and between the diffraction at the selected wavelengths. The anomalous scattering factor has a real part (Δf') and an imaginary part (Δf"). Usually, diffraction data are collected at three wavelengths (Fig. 2.4):

(i) First wavelength, where Δf ' has its maximum and where the Bijvoet differences are largest (also called 'peak', P).

(ii) Second wavelength, where Δf ' has its minimum (also called as 'edge', E).

(iii) Third wavelength, high (H) and/or low energy remotes (L) from the edge where Δf ' and Δf " are small.

Data collection: MAD data sets for TGEV M^{pro} were collected around the Se K-edge (12671.32 eV / 0.978470 Å) using a Mar165 CCD detector at beamline BW7A of the EMBL Outstation at DESY (Hamburg, Germany). Crystals of the SeMet-derivatized TGEV M^{pro} were isomorphous to the native ones except one crystal. Data were collected including an inverse beam sweep (ie. a sweep of data collected with the crystal rotated by 180º with

respect to the initial sweep) for the f " maximum and f ' minimum wavelengths, in order to guarantee the collection of all possible anomalous pairs.

Three different complete MAD experiments were done on three different SeMet-substituted crystals. The first crystal was non-isomorphous, so the data from this crystal could not be merged with those obtained from the other crystals, to get high redundancy. A four-wavelength experiment was conducted on the second crystal, whereas three-wavelength data were collected on the third isomorphous SeMet-derivatized protein crystal. With one, out of these two crystals, the peak-wavelength data were collected twice. Data were collected up to 2.8 Å resolution, though they were useful only up to 2.9 Å resolution (see Table 3.4).

Data integration and scaling were performed using the programs DENZO and SCALEPACK (Otwinowski & Minor, 1997). Self-rotation function calculations, performed using the CCP4 suite program ALMN (Collaborative Computational Project Number 4, 1994), suggested the presence of twofold non-crystallographic symmetry (NCS) axes. They were not used for determination of the Se–substructure due to their complexity. In retrospect, four independent NCS twofold axes can be found in the asymmetric unit, three within the dimers and one relating two dimers (A-B and E-F) (Section 3.2.5.2).

***HCoV M^{pro}*:** Prior to the MAD experiment with the SeMet TGEV M^{pro} crystals, Δ301-302 mutant SeMet-derivatized crystals were also tried for a MAD experiment at ELETTRA, Trieste, Italy. However, the selenium signal was too weak to solve the structure by direct methods. The data sets were collected at four wavelengths. All of them were merged together to obtain highly redundant data. This approach was useful for solving the structure by molecular replacement, taking TGEV M^{pro} as a model. Data integration and scaling were performed using the programs DENZO and SCALEPACK (Otwinowski & Minor, 1997). Self-rotation function calculations were performed using the CCP4 suite program ALMN (Collaborative Computational Project Number 4, 1994).

2.2.11 Structure determination

***TGEV M^{pro}*:** The bottleneck of the structure determination was the identification of the 60 selenium positions (6 monomers/a.u., with 10 Se each, were expected). The selenium substructure was determined with the *SnB* v2.0 (Weeks & Miller, 1999) package. *SnB* is a computer program based on a dual-space direct-methods procedure for determining crystal structures from X-ray diffraction data. It provides a graphical user interface for (i) computing normalized structure-factor amplitudes, (ii) the main phasing algorithm – Shake & Bake (Sheldrick, 1990a) and (iii) visualization and molecule-editing facilities (http://www.hwi.buffalo.edu/SnB).

Normalized difference structure factors (diffE) (Blessing & Smith, 1999) values were computed from the observed anomalous differences using the DREAR suite (Blessing *et al.*, 1996). All the data sets were scaled together into a single scale file. This file was then used in the program *SnB* v 2.0 to find out the positions of selenium atoms. Initially, only the peak wavelength (P1) data set from one crystal was used, but it was not enough to obtain a solution by *SnB* v2.0 (Weeks & Miller, 1999) (see Table 3.5). The redundancy of that data set was only 3.8. By increasing the redundancy through merging of various data sets increased the precision of the averaged intensities (as described by R_{pim}; Weiss & Hilgenfeld, 1997) sufficiently. In this way, the typical bimodal distribution was obtained in the *SnB* histograms indicative of some correct solutions. Taking only two high-energy remote (H1+H2) data sets gave 1 solution out of 4000 trials. Two merged peak-wavelength data sets (P1+P2) yielded 3 solutions out of 3000 trials; three merged peak-wavelength data sets (P1+P2+P3) gave 77 *SnB* solutions out of 2300 trials. Finally, all three peaks (P1+P2+P3), and two edge (E1+E2) wavelength data sets (redundancy = 18; Table 3.5) were merged together and then their anomalous differences were processed with the DREAR suite (Blessing *et al.*, 1996) to generate normalized difference structure factors (diffE values), giving 105 solutions from 5000 trials. The 1900 target diffE values were used to generate 20,000 triplet invariants. Between 2300 and 5000 trials were carried out with 120 *SnB* cycles of dual-space refinement on single data sets as well as merged data sets (see Table 3.5). For heavy-atom refinement and phase calculation at 2.9 Å resolution, all data sets were merged into a single "mtz" file using the CCP4 package (Collaborative Computational Project

Number 4, 1994). This file was then used in the program MLPHARE (Otwinowski, 1991) for the further refinement of these positions. After refining the positions, solvent flattening and NCS averaging was performed and an averaged mask of one of the monomers was obtained using the program DM (Cowtan & Main, 1996) that gave interpretable electron density maps. Subsequently, a mask was generated using the program MAPMASK (CCP4 suite) and NCS matrices were calculated from the selenium positions in each monomer. These matrices were improved after each refinement cycle. Full NCS averaging was performed with DM. The initial correlation coefficient between the monomers was about 0.48 in DM.

***NCS*:** Out of the 60-Se *SnB* solutions, 37 positions obeyed the NCS. A further 11 positions were predicted by the NCS, adding to a total of 48 out of 60 Se positions. Non-crystallographic symmetry restraints were applied during the initial stages of refinement against the 2.9 Å data; these restraints were gradually released as the resolution limit was extended to 1.96 Å. NCS averaging was used to generate meaningful estimates for missing reflections.

***HCoV M^{pro}*:** The structure was solved by molecular replacement (Rossmann, 1990) with the software package AmoRe (Navaza, 1994). The refined structure of TGEV M^{pro} was used as the search model after all residues had been replaced by their counterparts in HCoV M^{pro} and bad contacts eliminated. Crystallographic refinement was carried out with the program CNS (Brünger *et al.*, 1998a) and alternated with cycles of manual inspection and rebuilding. The refinement consisted of cycles of simulated annealing (Brünger *et al.*, 1990) followed by conjugate gradient minimization of atomic coordinates and isotropic temperature factors.

2.2.12 Indicators of diffraction data quality

The quality of the X-ray data is routinely assessed by different criteria. One of them is the symmetry (R_{sym}) or merging R factor that arises from the averaging of multiple measurements of reflections of the same (h,k,l) and of symmetry-related reflections. The second quantity is the ratio of the recorded intensity and its standard deviation, I/σ(I), and the third one is the redundancy of the data, *i.e.*, how often a given reflection and/or one of its symmetry-related reflections have been observed. The fourth quantity is the completeness of the data set, overall and in the highest resolution bin. Of these, R_{merge} is inherently dependent

on the redundancy of the data. The more often a given reflection is observed, the higher R_{merge} will be, even though by simple statistical reasoning the average value of the measurements will become more accurate (Weiss and Hilgenfeld, 1997; Diederichs & Karplus, 1997; Weiss, 2001). To overcome these drawbacks, a redundancy-independent merging R-factor (R_{rim}) and the precision-indicating merging R-factor (R_{pim}) have been proposed (Weiss and Hilgenfeld, 1997; Weiss, 2001). R_{rim} contains the redundancy N or the multiplicity of the observed reflection and is basically the conventional R_{merge} made independent of redundancy. R_{pim} also contains the redundancy N and indicates how precisely the average measurement has been measured.

$$Rmerge = \sum_{h,k,l} \sum_{i} | I_i(h,k,l) - I(\bar{h},\bar{k},\bar{l}) | / \sum_{h,k,l} \sum_{i} I_i(h,k,l)$$

$$Rrim = \sum_{h,k,l} \sqrt{\frac{N}{N-1}} \sum_{i} | I_i(h,k,l) - I(\bar{h},\bar{k},\bar{l}) | / \sum_{h,k,l} \sum_{i} I_i(h,k,l)$$

$$= R_{mean} \text{ (Diederichs \& Karplus, 1997)}$$

$$Rpim = \sum_{h,k,l} \sqrt{\frac{1}{N-1}} \sum_{i} | I_i(h,k,l) - I(\bar{h},\bar{k},\bar{l}) | / \sum_{h,k,l} \sum_{i} I_i(h,k,l)$$

Where I_i is the observed intensity and <I> is the average intensity from multiple measurements. *N* is the number of times a given reflection has been measured. R_{rim} corresponds to an R_{sym} that is independent of the redundancy of the measurements.

In the case of TGEV M^{pro}, it was impossible to solve the structure with one wavelength data because of the low redundancy (3.8); also indicated by the worse R_{pim} value. More data sets had to merge together to gain as high redundancy as possible (18.0-fold redundancy was achieved) (for details see Table 3.5). The same was done for the HCoV M^{pro} diffraction data sets. All four wavelengths (the MAD experiment failed to solve the phase

problem because of very poor selenium signal) were merged together to get a highly redundant data set to solve the phase problem by the molecular replacement method.

2.2.13 Model building and refinement of TGEV M^{pro} and HCoV M^{pro}

Interpretable electron density maps are essential for guiding a proper model building. This is facilitated by drawing contour maps at various density levels. A higher electron density (ρ) at a given location (x,y,z) indicates the probable occurrence of an atom. Model building entails that these high-density locations are fitted with appropriate atoms. As the model building proceeds, it is imperative to ascertain the correctness of the built model at regular intervals. This is done in two ways by calculating the difference between the structure factor values of the observed (F_o) and calculated (F_c) maps.

$$(Fo-Fc)map:\rho(x,y,z)=\frac{1}{V}\sum_{h,k,l}(|Fo|-|Fc|)e^{-2\pi i(hx+ky+lz-\alpha_{calc})}$$

The map indicates where the model should be adjusted to increase the electron density in this region by moving atoms towards that location, and *vice versa* for the negative density regions.

$$(2Fo-Fc)map:\rho(x,y,z)=\frac{1}{V}\sum_{h,k,l}(2|Fo|-|Fc|)e^{-2\pi i(hx+ky+lz-\alpha_{calc})}$$

α= phase, V = unit cell volume

The 2Fo-Fc map shows the difference between the actual structure and the model in addition to the observed electron density of the model. In this map, the model influence is reduced, but not as severely as in the Fo-Fc map.

The initial electron density maps for both TGEV M^{pro} and HCoV M^{pro} obtained from preliminary phases were interpretable. For both the proteinases, initially the A monomer was built manually into the electron density map, then the other monomers were generated at low resolution by NCS symmetry operation(s). The TGEV M^{pro} had 6-fold NCS, which was applied at 2.9 Å resolution. The monomer B of HCoV M^{pro} was generated by two-fold NCS operation in a similar manner. A polyalanine model of 90% of a single subunit was built with the program 'O' (Jones *et al*., 1991). Most of the side chains were identified afterwards. The known locations of the Se sites allowed the assignment and direction of the peptide chain.

***Omit map*:** An important way to overcome phase bias is the use of omit maps. An 'OMITMAP' is made by removing the residues of interest from the model for calculating the phases. In theory, this will allow the phases calculated from the rest of the model to phase the area of interest with no bias from the model left out. The method takes advantage of the Fourier transform property that every point in real space is influenced by every point in reciprocal space, and *vice versa*. In TGEV M^{pro}, the flexible surface loop from residues 216 to 225, the C-terminus of helix E, the loop region between residues 267 and 276, and the segment 294-300 following the C-terminal F helix were not having good electron density. Likewise the loop region 46-49; and the region 242-259 of HCoV M^{pro} were poorly defined. Many cycles of the 'OMITMAP' procedure finally resulted in good electron density map for these regions.

2.2.13.1 Introduction of water molecules into the structure

A significant part of the model are ordered water molecules. The following criteria were employed in the introduction of water molecules in the structure.

They should represent the residual (Fo-Fc) electron density 4σ above the mean level in the electron density map and 1σ above the mean in the 2Fo-Fc map. More importantly they should have chemically reasonable distances to potential hydrogen-bond donors/acceptors. The improvement of the model was monitored by both the conventional R-factor and R_{free} (Brünger, 1992a). The B-factor and coordinates for the water molecules were refined in the water-picking script of the program CNS.

2.2.13.2 Refinement of the TGEV proteinase structure to high resolution

Initial rounds of refinement involved a combination of CNS and Arp/warp (Lamzin & Wilson, 1997) for the high resolution (1.96 Å) data from TGEV M^{pro}. The refinement was carried out against 95% of the measured data. The remaining 5% that were randomly excluded from the full dataset were used for cross-validation by calculating the free R-factor (R_{free}) to follow the progress of refinement (Brünger, 1992a). The same set of reference reflections was used throughout the refinement. The reference set was also excluded from the calculation of the electron density maps.

The refinement was performed with CNS (Brünger *et al*., 1998b). The procedure included simulated annealing, B-factor and conjugate gradient energy minimization against maximum likelihood targets as implemented in the program CNS (Brünger *et al*., 1998a). After each step, 2Fo-Fc and Fo-Fc electron density maps were calculated and the model was visualized and rebuilt using the program 'O' (Jones *et al*., 1991). Rebuilding proceeded by systematically checking all the electron density peaks greater than 4σ in the Fo-Fc Fourier maps and building the missing residues which were removed during the beginning of refinement; the σ cut-off was systematically lowered during the later cycles of refinement.

When the resolution was extended to 1.95 Å, it was found that there were some pieces in each of the monomers which were not NCS-related. So later stages of refinement cycles were independent of NCS. Final refinement converged to free R factor of 25.6% and a crystallographic R-factor of 21.0%.

The crystal structure refinement of HCoV M^{pro} was done in quite a similar manner as for TGEV M^{pro} structure. The refinement converged to a free R factor of 28.8% and a crystallographic R factor of 22.2%.

2.2.14 Assessment/Validation

***The quality and structure analysis of TGEV M^{pro} and HCoV M^{pro}*:** All the monomers in the crystal structure of TGEV and HCoV M^{pro} were compared using the programs ALIGN (Cohen, 1997) and LSQKAB (CCP4, 1994). Visual comparisons were made using program 'O' (Jones *et al.*, 1991). Analysis of the crystal contacts was performed using the program CONTACT from the CCP4 Suite (1994).

The final models of all structures were validated by the program PROCHECK (Laskowski *et al*., 1993) to check the overall quality using the Ramachandran plot and other stereochemical criteria. The errors in the atomic coordinates of the molecular model were estimated using a Luzzati plot (Luzzati, 1952). The quality of macromolecular structure-factor data and their agreement with the atomic model was checked with the program

SFCHECK (Vaguine *et al*., 1999). This indicator takes into account errors in the data, atomic displacement factors, quality of the crystal and series termination effects (Blundell & Johnson, 1976) and indicates a correlation factor between the observed and the calculated structure factor amplitudes. It is relatively independent of the choice of the nominal resolution of the data set and should therefore be objective. It is also relatively independent of the completeness of the data set. Further inconsistencies were identified using the program WHATCHECK (Vriend, 1990). This program checks the quality of protein structures for many different kinds of errors. It produces an elaborate report, ranging from trivial bond length, torsion angle, and surface checks to highly advanced contact analysis and hydrogen-bonding network checks. The solvent accessibility of the monomers was calculated using the algorithm of Lee and Richards (1971) as implemented in the program NACCESS (http://sjh.bi.umist.ac.uk/naccess.html). A solvent probe of radius 1.4 Å was used for the accessibility calculations. The inter-subunit contacts and the contacts between the symmetry mates were probed by counting the atoms in a sphere of radius 3.8 Å. The electrostatic potential surface of the monomer was calculated using the program GRASP (Nicolls *et al*., 1991). The molecular diagrams were drawn using INSIGHT II (Biosym/MSI, San Diego, California, USA, 1995), Molscript (Kraulis, 1991) and Bobscript (Esnouf, 1999). The diagrams were rendered using the program Raster3D (Bacon & Anderson, 1988; Merritt & Bacon, 1997). Molscript, Bobscript and Raster3D are the programs of choice to draw the three-dimensional models highlighting the secondary structures. Bobscript can additionally draw electron density maps around the desired residues.

2.2.15 Sequence analysis and three-dimensional structure search

Primary structure analysis was carried out with the GCG software suite (Wisconsin package version 9.0, Genetics Computer Group (GCG), Madison, USA). Sequence motifs were explored using all the domains or by taking only domain III, and taking search patterns as defined by the PROSITE dictionary of protein sites and patterns. The PDB files were searched for folds similar to domain III using the programs DALI (Holm & Sander, 1993) and TOPS (Gilbert *et al.*, 1999).

2.3 TGEV M^{pro} in complex with TLCK

The amount of protein in the crystal was calculated as 3.48 x 10^{-10} Moles. This was done assuming 12 copies of the protein per unit cell (72.8x160.1x88.8 $Å^3$) and crystal dimensions of (300x300x200) x104 $Å^3$. Accordingly, one crystal was soaked overnight in 30 µl of 1mM TLCK (Sigma, Germany). The data were collected up to 2.6 Å resolution at using the Joint IMB Jena-University of Hamburg-EMBL synchrotron beamline X13 at Deutsches Elektronen-Synchrotron, Hamburg, at a wavelength of 0.802 Å at crystal temperature of 100 K. A diffraction image from the X-ray data collection is shown in Figure 2.5.

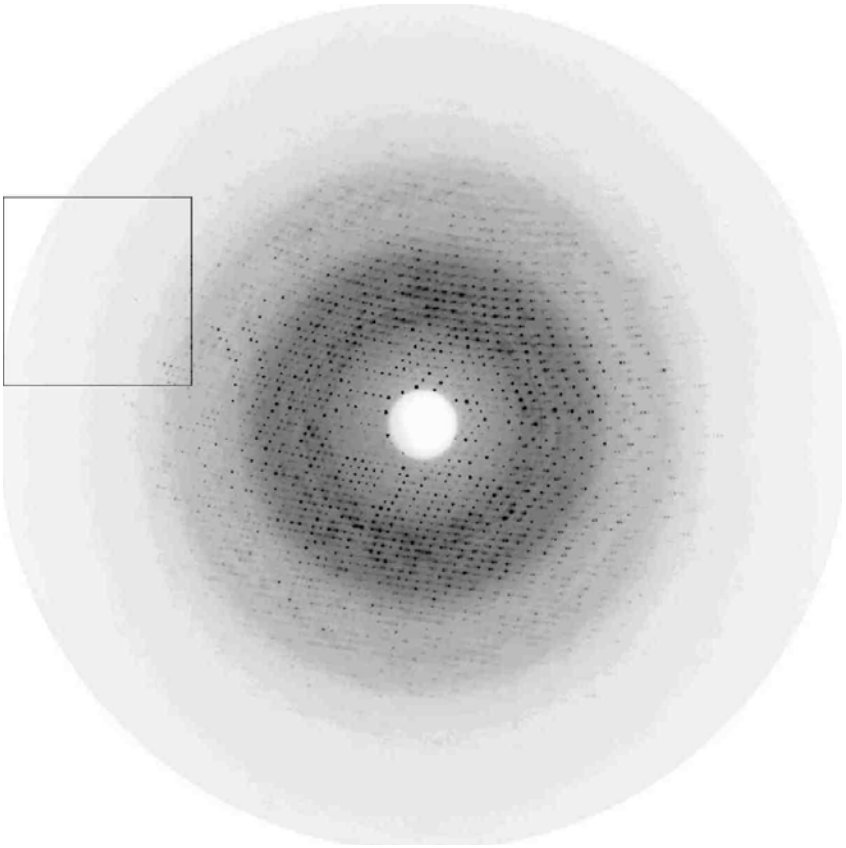


Fig. 2.5 Sample diffraction image of TGEV M^{pro}-TLCK crystal. Data collected using the Joint IMB Jena-University of Hamburg-EMBL synchrotron beamline X13 at Deutsches Elektronen-Synchrotron, Hamburg,

2.3.1 TGEV M^{pro}-TLCK Structure

Soaking with TLCK did not change the unit-cell dimensions of the TGEV M^{pro}-TLCK complex crystal. The starting phases were obtained directly from the TGEV M^{pro}, from which water molecules and other ligands were removed.

2.3.2 Refinement of the TGEV M^{pro}-TLCK structure

The refinement cycles were performed using the program CNS (Brünger *et al*., 1998a). All model building and graphical manipulations were performed with the program 'O' (Jones *et al*., 1991).

2.4 TGEV M^{pro}-CMK inhibitor complex (substrate analog)

2.4.1 CMK-Hexapeptide synthesis and Purification: Dr. Parvesh Wadhwani (Institute für Molekular biologie, Jena, provided the synthesized, purified and analyzed peptide. The peptide synthesis was performed using an Applied Biosystem's 433A peptide synthesizer. Conversion of the free C-terminal to the chloromethyketone functionality was performed as previously reported (Krantz *et al*., 1991).

2.4.2 Soaking, data collection and refinement: TGEV M^{pro} crystals were soaked with fivefold molar excess of CMK inhibitor. Crystals were soaked for about 16 hr, quickly rinsed into the mustard oil at 4^{o}C for cryo-protection then were taken for the measurement. A full data set to 2.2 Å resolution was subsequently collected using the Joint IMB Jena-University of Hamburg-EMBL synchrotron beamline X13 at Deutsches Elektronen-Synchrotron, Hamburg, at a wavelength of 0.802 Å, at crystal temperature of 100 K. The data were processed using DENZO and SCALEPACK programs (Otwinowski and Minor, 1997). The structure was refined in many cycles using program CNS (Brünger *et al*., 1998a). All model building and graphical manipulation were performed with the program 'O' (Jones *et al*., 1991; Jones and Kjeldgaard 1995).

3 RESULTS AND DISCUSSION

3.1 Crystallization of recombinant coronavirus main proteinases

TGEV M^{pro}: The protein solution containing the storage buffer (see Table 2.3) was screened for potential crystallization conditions using the commercially available kits from Hampton Research. Screening was done at 4, 10, 15, 20 and 30 ^{o}C using different pH ranges. The experiments were performed using the vapor diffusion method (hanging drop) (Fig. 2.2). Purified protein solution containing storage buffer (Section 2.2.1) was mixed with an equal volume (1µl) of precipitant solution and equilibrated over reservoir containing 1 ml of the latter. Only trials at 4 ^{o}C and at alkaline pH (pH 8.8) gave any crystals for TGEV M^{pro}. The microcrystals appeared after many screening rounds. The quality was improved upon including an additive like dioxane. Larger pyramidal TGEV M^{pro} crystals were finally obtained using a protocol with optimized conditions (1.8 M ammonium sulfate, 6% MPD, 100 mM HEPES pH 8.8, 5 mM DTT). 4% dioxane was added later to the reservoir solution only, after setting up the drop. It was thus not directly added to the protein drop but was present in the equilibration process. The crystals obtained are displayed in Figure 3.1A. SeMet-derivatized crystals were obtained under similar conditions except for 2 M ammonium sulfate and 8% MPD (Fig. 3.1B).

HCoV M^{pro}: Initial rounds of screening gave clusters of very thin needles for HCoV M^{pro}. Attempts to improve the size and quality by growing crystals in the presence of various additives resulted in large plate-like, diffractable crystals of SeMet-derivatized HCoV M^{pro}, after using optimized conditions (20% PEG 10K, 2% 1,6-hexanediol and 12% dioxane, at 10 ^{o}C, 100 mM HEPES pH 8.5) (Fig. 2.3A, B). Numerous problems were encountered during the crystallization. Successful crystallization was batch-dependent and even crystallizable batches exhibited poor reproducibility. Furthermore, these crystals predominantly displayed severe defects including stacking of many plates on one another, cracking, and splitting, branching and satellite formations. The optimized crystallization conditions for TGEV and HCoV M^{pro}s are given in the Table 3.1.

Table 3.1 Optimized crystallization protocols

HCoV M^{pro}	15% PEG 6K, 4% 1,6-hexanediol, 6 mM DTT, 0.5% ethanol, 100 mM HEPES, pH 8.5, 10 ^{o}C*
SeMet-HCoV M^{pro}	20% PEG 10K, 2% 1,6-hexanediol, 6 mM DTT, 100 mM HEPES, pH 8.5, 10 ^{o}C*
TGEV M^{pro}	1.8 M ammonium sulfate, 6% MPD, 5 mM DTT, 4% (v/v) dioxane, 100 mM HEPES pH 8.8, 4 ^{o}C*
SeMet-TGEV M^{pro}	2 M ammonium sulfate, 8% MPD, 5 mM DTT, 4% (v/v) dioxane, 100 mM HEPES pH 8.8, 4 ^{o}C*

* Temperature

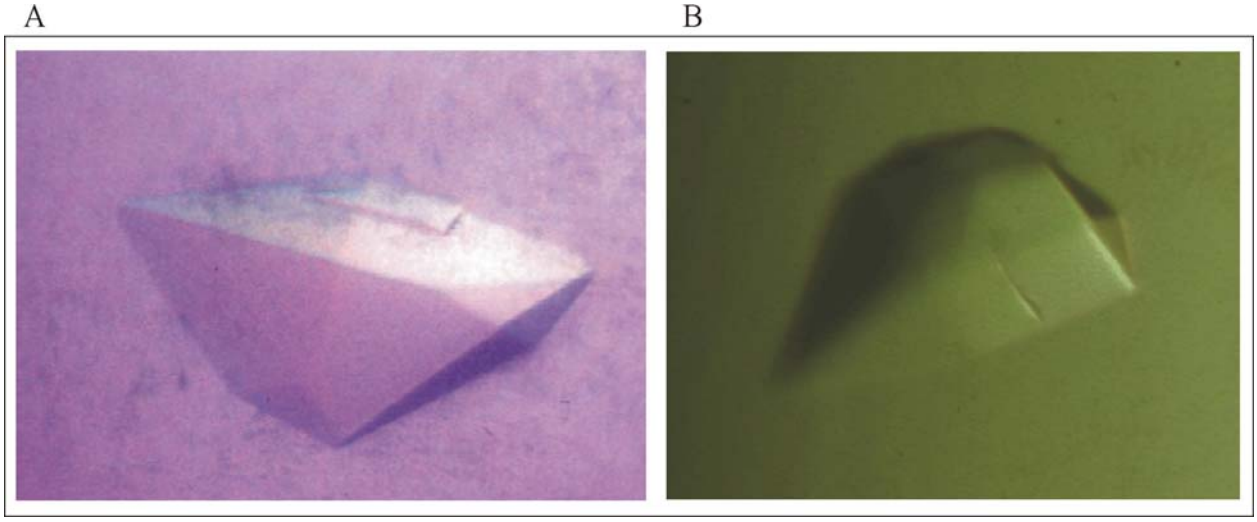


Fig. 3.1 **A.** Native crystal of TGEV M^{pro} grown by the hanging drop method. The approximate dimensions of these crystals are 0.3 x 0.25 x 0.3 mm^3. **B.** SeMet-derivatized crystal of TGEV M^{pro}.

3.2 Structure elucidation

3.2.1 Native-data acquisiton

The TGEV M^{pro} crystals were fragile and suffered from radiation damage when exposed to X-rays for longer durations. As a result, collecting data at room temperature was difficult. Therefore, to cryo-cool the crystals, many conventional and non-conventional cryoprotectants were tried (Section 2.2.7), none of which worked. Finally, a quick rinse in crude mustard (*Brassica compestris)* oil (mustard seed oil, Delhi, India) (Fig. 3.2) was successful. Crystals were immediately cryo-cooled in liquid nitrogen. To overcome the increased mosaicity

problem, the mustard oil was kept at the same temperature as the crystals were grown, i.e. at 4 ^{o}C.

Fig. 3.2 The source of mustard oil is seeds from the mustard plant (*Brassica compestris)*.

Data sets from crystals of native TGEV M^{pro} diffracting up to 1.95 Å resolution, were collected at 100 K on the X-ray diffraction beamline at ELETTRA, Trieste, using a Mar165 CCD detector (Table 3.2).

The crystals of HCoV M^{pro} were extremely fragile, thin plates and prone to cracking upon handling. Therefore, the specimens could not be transferred to a storage buffer or cryogenic buffer without undergoing severe and irreparable damage. Fortunately, crystals could be cryo-cooled without any cryoprotectant as the 20% PEG 10K used in the crystallization medium acted as a cryoprotectant itself. Data from native-HCoV M^{pro} crystals were difficult to scale and had low redundancy, and were therefore discarded. Subsequently, data from the SeMet-derivatized crystals were used to solve the structure. MAD data collected at all four wavelengths were merged together (see section 3.2.5) to obtain highly redundant data set up to 2.6 Å resolution.

3.2.2 X-Ray diffraction data

***TGEV M^{pro}*:** The unit-cell dimensions (a = 72.81 Å, b = 160.13 Å, c = 88.96 Å, β = 94.34^{o}) as well as the self-rotation function implied that several monomers (4-8) were present in the asymmetric unit. The high-resolution native data set obtained at the synchrotron source gave diffraction data with an average value of I/σ(I) of 28.0 for all reflections (resolution range 50-1.96Å) and 3.98 in the highest resolution shell (1.98-1.96 Å). The data set was more than 98% complete. A total of 787,522 measurements were made, representing 144,735 independent reflections. Data processing gave R_{merge} 4.2%, R_{pim} = 1.8%, R_{rim} = 4.6%, and a mosaicity of 0.6^{o}. The Matthews coefficient for 6 molecules per asymmetric unit was 2.53 Å^3/Da and the solvent content was 51.4% (Matthews, 1968). A summary of the X-ray diffraction data is given in Table 3.2.

***HCoV M^{pro}*:** Crystals displayed space group $P2_1$ with unit cell dimensions a = 53.3 Å, b = 76.1 Å, c = 73.4 Å, β = 103.7^{o}, and two proteinase monomers in the asymmetric unit. The data were collected at the synchrotron source equipped with a Mar 345 detector. Diffraction data collected gave average value for I/σ(I) of 16.44 for all reflections (resolution range 25-2.5 Å) and 9.1 in the highest resolution shell (2.59-2.50 Å). The data were more than 98% complete. A total of 228,513 measurements were made, representing 17,533 independent reflections. Data processing gave R_{merge} = 14.2%, R_{rim} = 14.2%, R_{pim} = 3.0%, and a mosaicity of 1.8^{o}. The Matthews coefficient for 2 molecules per asymmetric unit was 2.21 Å^3/Da and the solvent content was 44.4% (Matthews, 1968). Diffraction data are summarized in Table 3.2.

Table 3.2 Summary of X-ray diffraction data: native crystals

Diffraction data statistics	**TGEV M^{pro}**	**HCoV M^{pro}**
X-ray source	Synchrotron radiation[a]	Synchrotron radiation[a]
Detector	Mar CCD	Mar 345
No. of frames	540	600
Crystal oscillation (^{o})	0.5	1.0
Wavelength (Å)	0.99983	0.980122
Temperature (K)	100	100
Space group	$P2_1$	$P2_1$
Unit cell parameters (Å, ^{o})	a = 72.81, b = 160.13, c = 88.96, β = 94.34	a = 53.3, b = 76.1, c = 73.4, β = 103.7
Resolution (Å)[b]	50.0-1.96 (1.99-1.96)	25-2.60 (2.69-2.60)
Completeness (%)	98.2	98.9
R_{merge}(%)[b,c]	4.2 (22.1)	14.2 (41.2)
R_{rim} (%)[b,c]	4.6 (27.1)	14.2 (43.9)
R_{pim} (%)[b,c]	1.8 (15.2)	3.0 (13.0)
Redundancy	5.4	12.3
I/σ(I)	3.4	9.1
Mosaicity (^{o})	0.62	1.8
No. of reflections measured	787,522	216,984
Unique reflections	144,735	17,533

[a] X-ray diffraction beamline at ELETTRA, Trieste, equipped with a Mar CCD detector at the time of HCoV M^{pro} data collection it was equipped with Mar 345.

[b] Highest resolution bin in parentheses

[c] For definition, see Section 2.2.12

3.2.3 Initial attempts to structure solution

***HCoV M^{pro}*:** Initial attempts to solve the structure were made with HCoV M^{pro} crystals. The amino-acid sequence was submitted for a BLAST search (URL: http://www.ncbi.nlm.nih.gov/BLAST) to identify promising homologous three-dimensional models, which resulted in 152 hits. The best hit had a score of 60% sequence identity with its own family member, however, for none of the family members, there was a three-dimensional model in the PDB archive. The rest of the results were with sequence identity below 15% so

were not suitable templates for search models. Additionally, these structures did not have folds similar to the one predicted by the secondary structure prediction program PHD (Rost & Sander, 1993) or a β-barrel fold as predicted by Gorbalenya *et al*. (1989b). Therefore, $3C^{pro}$ structures were downloaded from the PDB for three-dimensional superposition of their polypeptide backbones. These structures were input into the molecular replacement programs AmoRe (Navaza, 1994), Mol-Rep (CCP4 suite, 1994) and EPMR (CCP4 suite, 1994) to solve the phase problem. Unfortunately, none of the models taken individually or in combinations in a superimposed form gave any suitable template, which could give any useful phase information. Soaking and co-crystallization with heavy-atom compounds was also tried without any success because of poor binding of the metal ions, compounded by the poor quality of the crystals. The next trial was therefore made on the TGEV M^{pro} crystals, which were of better quality and diffracting to higher resolution than HCoV M^{pro}.

***TGEV M^{pro}*:** Due to the low sequence identity to other proteinases and complicated by the fact that several molecules per asymmetric unit were expected; the structure could not be solved using conventional molecular replacement techniques**.** Thus, phasing methods such as multiple isomorphous replacement (MIR), and single or multiple isomorphous replacement with anomalous scattering contributions (SIRAS and MIRAS, respectively), were required. All these techniques need suitable heavy-atom complexes, prompting heavy-atom soaking and screening experiments. Different heavy-atom solutions of varying concentrations were added directly to the crystallization drop, following growth of suitable native crystals. Crystals were allowed to be soaked for various lengths of time prior to cryo-cooling and subsequently, analyzed for potential heavy-atom attachment. Unfortunately, none of the 35 heavy-atom compounds tried for TGEV M^{pro} successfully bound to it. To make the search productive and timesaving, native polyacrylamide gel electrophoresis (PAGE) of protein and heavy atom mixtures was used to search for gel shifts upon derivatization. This technique (Boggon & Shapiro, 2000) can show which reagents cause protein denaturation and are therefore less likely to be useful as heavy atom derivatives. There was a band shift observed for phenyl mercuric acetate, $K_2Pt(CN)_4$, and cis-platine II (Fig. 3.3). The band shift observed in these cases resulted from retardation of mobility, but the analysis of the diffraction data showed that even these compounds were bound only partially. Additionally, there was a significant non-isomorphism between different crystals. All the native and heavy-atom-derivatized crystals were measured at similar temperature, i.e. 100 K. It was not possible to

scale data sets from different crystals together. Consequently, multiwavelength anomalous diffraction (MAD) experiments seemed most suitable for solving the phase problem.

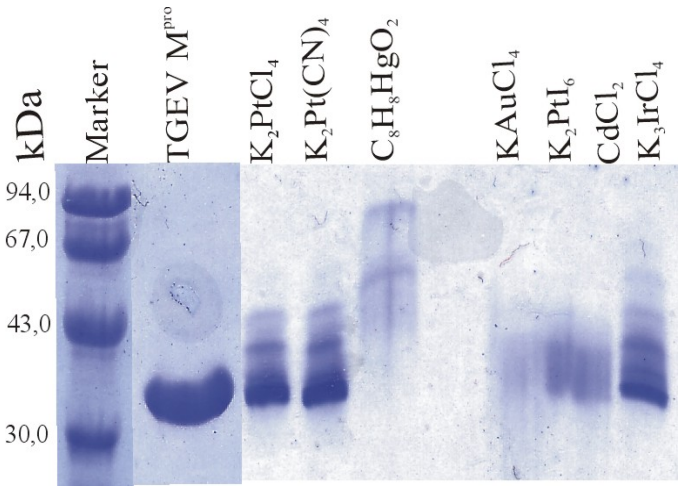


Fig. 3.3 Native gel showing band shifting when heavy-atom compound bound to the TGEV M^{pro}.

Consequently the next recourse was expression and purification of SeMet-derivatized M^{pro}s, which was done by Dr. Ziebuhr (Würzburg). First, the crystals of SeMet-HCoV M^{pro} were tried, but both the peaks for the mass of SeMet-derivatized protein in the MALD-TOF spectra (Fig.3.5) and the fluorescence peak of the SeMet-HCoV M^{pro} crystal were rather bleak. Additionally, the crystals and the data set were not of good quality, which persuaded to determine the TGEV M^{pro} structure first and the related structure of HCoV M^{pro} by molecular replacement later. The comparative details of the two derivatized proteins is described in the next section.

3.2.4 Characterization of coronavirus SeMet-derivatized M^{pro}s

***TGEV M^{pro}*:** Ziebuhr *et al* (1997a) found the molecule to be in the monomeric state in solution by gel filtration experiments. Figure 3.4 shows a typical spectrum of the protein as obtained by MALDI mass spectrometry (Section 2.2.4.1). The apparent mass was detected as 33.17 kDa. The calculated molecular mass derived from the predicted wild-type amino acid sequence is 33.09 kDa. Comparison of predicted mass to the mass measured by MALDI mass spectrometry indicates an error of 0.2%, which is within the error margin of the instrument.

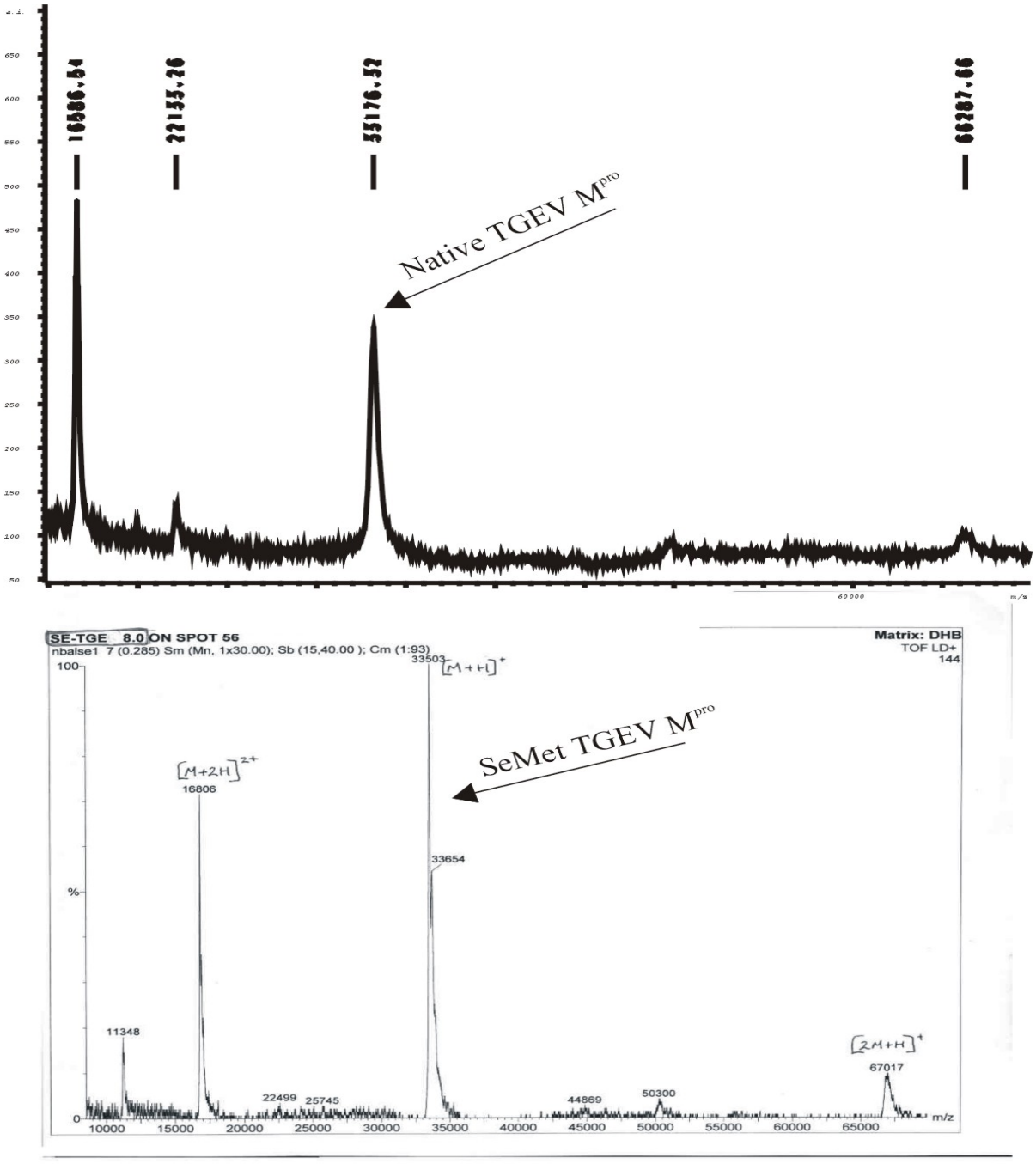


Fig. 3.4 Spectra obtained by MALDI-TOF for the native (upper) and SeMet-derivatized protein (lower) of TGEV M^{pro}.

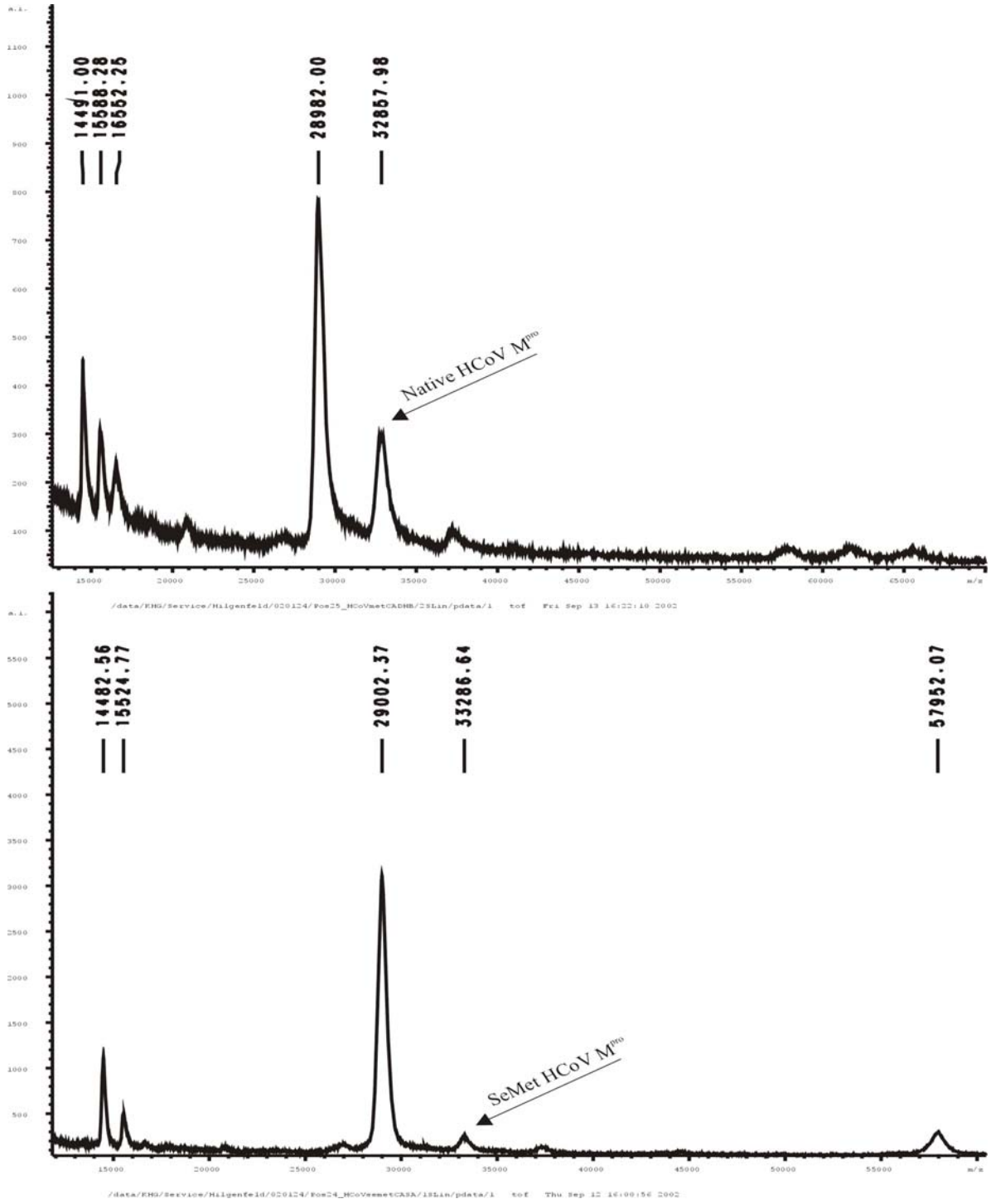


Fig. 3.5 Spectra showing the accurate molecular mass of the native (upper) and SeMet-derivatized protein (lower) of HCoV M^{pro}, obtained from MALDI-TOF.

The mass spectra for SeMet-derivatized TGEV M^{pro} showed a peak at 33.51 kDa. The difference of mass shown by MALD-TOF between the wild-type TGEV M^{pro} and SeMet-derivatized TGEV M^{pro} indicated about 70% of selenium incorporation in the protein, when calculated for 10 methionines. This was a good encouragement to crystallize this protein.

Preliminary dynamic light scattering experiments (Section 2.2.4.2) showed about 35% of the protein mass in the dimeric form. TGEV M^{pro} is a mixture of monomer (29.8 kDA) and dimer (67.5 kDa) in protein storage solution (11mM Tris.HCl pH 7.4, 120mM NaCl, 0.1mM EDTA, 1mM DTT – buffer A, Section 2.2.4.2). When repeated under crystallization condition, mixing buffer A and buffer B (1.8 M ammonium sulfate, 6% MPD, 100 mM HEPES pH 8.8), the experiment showed the dimer as by far dominant species. Under crystallization conditions, the higher order aggregates (~200 kDa) were seen (buffer B), which was a clear indication that protein is not monomeric under crystallization conditions. The results were consistent when repeated a number of times (Table 3.3). A dimeric arrangement with approximate C_2 symmetry is also found in the crystal. Like TGEV M^{pro}, the HCoV proteinase is also a homodimer in the crystal, though it was crystallized under completely different conditions and in a different unit cell.

Table 3.3 Dynamic light scattering data fot TGEV M^{pro}

Expt. No.	Protein buffer	Hydrodynamic radius (nm)	Polydispersity Intensity (%)	Estimated mass (kDa)
1.	Buffer A	2.54	32.9	29.8
2.	Buffer A	3.54	37.5	67.5
3.	Buffer B	4.17	02.2	206.7

***HCoV M^{pro}*:** The proteinase was found as monomer in solution by gel filtration experiments (Ziebuhr *et al*., 1997). MALDI-TOF spectra showed the mass of the C-terminal deletion mutant (Δ301-302) of HCoV M^{pro} determined by mass spectroscopy was 32.85 kDa, is within the error margin of the instrument (the calculated mass is 32.81 kDa) (Fig 3.5) and the mass of SeMet-derivatized M^{pro} was 33.2 kDa. The difference between both masses indicates about 70% selenium incorporation for 10 methionines. But the reliability of the calculations was diminished due to the weak signal for the derivatized protein (Fig 3.5, lower panel).

3.2.5 Data collection for SeMet-derivatized TGEV M^{pro} crystals

3.2.5.1 Multiwavelength anomalous dispersion (MAD)

MAD data sets for TGEV M^{pro} were collected around the Se K-edge using a Mar165 CCD detector at beamline BW7A of the EMBL Outstation at DESY (Hamburg, Germany) (Fig. 3.6). A sharp 'white line' feature was observed (Fig. 3.7). The number and position of points of measurement were chosen so as to optimize the strength of the signals in the experiment.

At the minimum, two wavelengths are needed for a definitive evaluation of the unknown heavy-

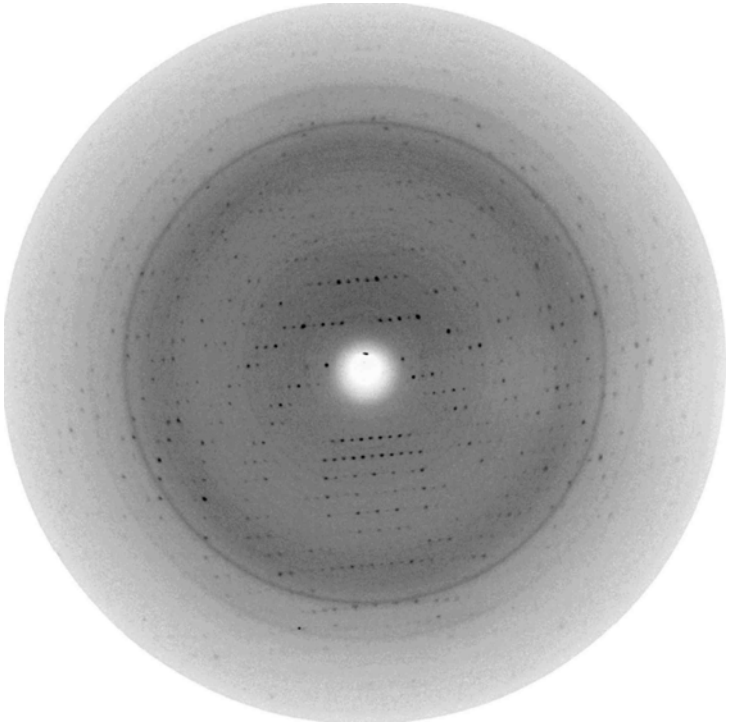


Fig. 3.6 Diffraction image of a SeMet-TGEV M^{pro} crystal (resolution 2.8 Å), measured at the BW7A beamline at EMBL, DESY, Hamburg. Although not apparent from the picture, diffraction spots of low intensity are present in the outermost rings. Crystal to detector distance is 205 mm.

atom position, but at least three are required to allow both the dispersive and Bijvoet differences to be optimized. Appropriate choices can be made from the experimental absorption spectrum. Clearly, diffraction data must be collected at an energy corresponding to the peak of absorption (f '') (Fig. 3.8). The second series of diffraction data must be collected at the edge inflection point f ' these are needed to maximize dispersive differences. When choosing the position for the high- and low-energy remote, it is important to note that systematic errors increase, as one gets further away from the edge. Most successful experiments (Table 3.4, 3.5) are carried out with the remote energies restricted to within 100 to 300 eV of the edge. Data were measured at 100 K from crystals that were flash-cooled with liquid nitrogen using mustard oil as a cryo-protectant.

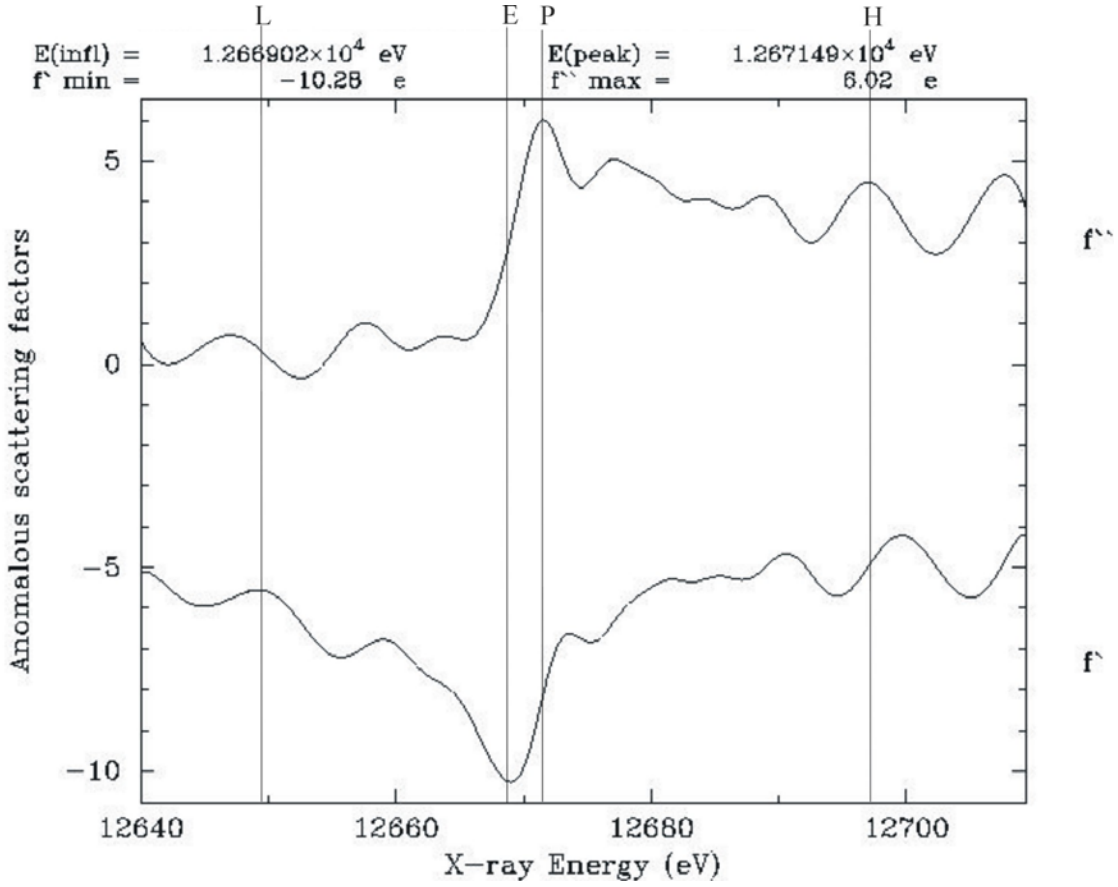


Fig. 3.7 Plot of f ' and f " calculated from a fluorescence scan on a SeMet-TGEV M^{pro} crystal, showing the data measured at low energy remote, f " maximum, f ' minimum and high energy remote.

The minimum values for f ', obtained from the inflection points (E1, E2) of the fluorescence spectrum, were determined to be 12.6678 keV (0.97874 Å) and 12.6690 keV (0.97864 Å). The peak (P1, P2, P3) wavelengths from two crystals were determined to be 12.6714 (0.97846 Å), 12.6715 (0.97845 Å) and 12.6712 (0.97848 Å).

A reasonable Matthews coefficient (Matthews 1968) of 2.3 $Å^3$/Da and a solvent content of 51% was obtained assuming six molecules in the asymmetric unit. The SeMet-TGEV M^{pro} data were consistent with the monoclinic crystal system ($P2_1$), with unit cell parameters a = 72.81, b = 160.12, c = 88.95 Å and β = 94.3°. Additionally, self-rotation function calculations (CCP4 suite) suggested the presence of noncrystallographic twofold rotation axes.

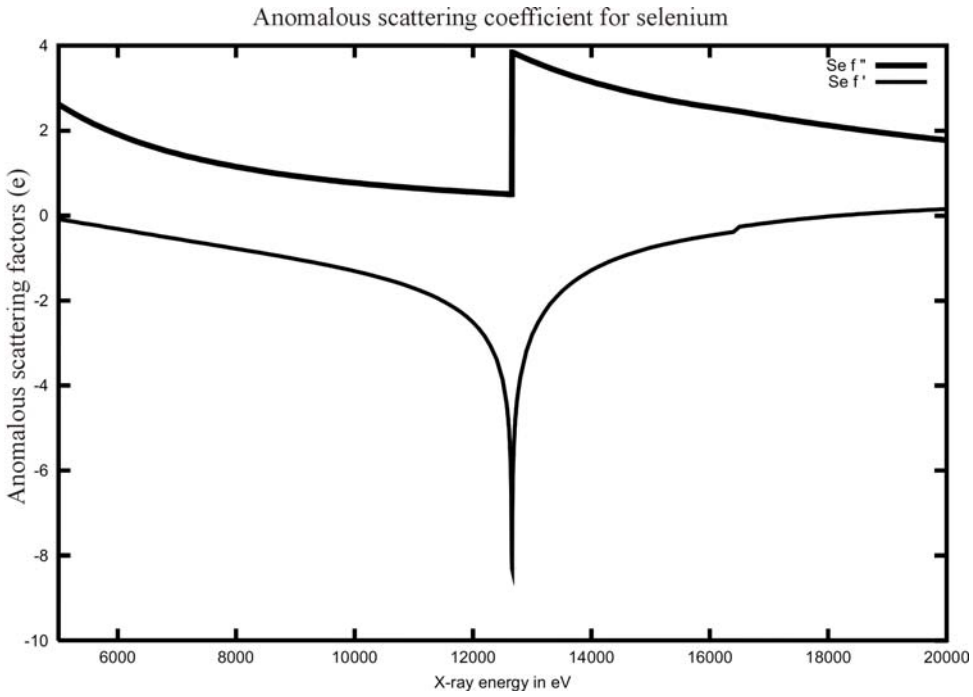


Fig. 3.8 Theoretical plots for f ' and f " over the K absorption edge of selenium. Data for this plot were obtained from Merritt (1998) (http://www.bmsc.washington.edu/scatter/As_form.html). The theoretical absorption edges occur at 12.6580 keV (K-II /E/edge).

Therefore, 60 crystallographically independent selenium sites had to be located as major part of the MAD phasing process.

The majority of the crystals of the SeMet-derivatized TGEV M^{pro} were isomorphous to the native ones except one (therefore, the data from this crystal could not be merged with those obtained from the other crystals). Data were collected including an inverse beam sweep (ie. a sweep of data collected with the crystal rotated by 180º with respect to the initial sweep) for the Δf " maximum and Δf ' minimum wavelengths, in order to guarantee the collection of all possible anomalous pairs. A four-wavelength experiment was conducted for one crystal, whereas three-wavelengths data were collected on other isomorphous crystals. A data set collected from a native crystal was used as low-energy remote for the second crystal. All the data from SeMet-crystals were used up to 2.8 – 2.9 Å resolution. A diffraction image taken of a SeMet-TGEV M^{pro} crystal is shown in Figure 3.6. Three different MAD data sets were collected using three different SeMet-substituted crystals. With one of the crystals, the peak wavelength was measured twice. A summary of the MAD experiment data is given in Table 3.4.

Table 3.4 Summary of X-ray diffraction data (SeMet substituted TGEV M^{pro})

Beamline	BW7A[a]							
Data set[b]	**P1**	**P2**	**P3**	**E1**	**E2**	**H1**	**H2**	**L1**
Wavelength (Å)[c]	0.97487	0.97845	0.97848	0.97864	0.97874	0.95583	0.9080	1.0022
Resolution (Å)	30-2.8	30-2.8	30-2.8	30-2.8	30-2.8	30-2.8	30-2.8	30-2.8
Completeness (%)	99.9	98.1	99.7	99.9	99.7	99.7	98.8	97.3
Mosaicity (°)	0.4	0.6	0.7	0.4	0.6	0.4	0.6	0.4
R_{merge} (%)[d]	10.5	11.4	10.6	8.1	8.2	8.6	7.2	8.0
R_{rim} (%)[e]	12.1	13.0	12.3	9.2	8.9	10.2	7.5	10.3
R_{pim} (%)[f]	6.1	6.6	6.4	4.7	4.5	5.2	3 .2	5.4
Redundancy	3.8	3.8	3.9	3.8	3.9	3.7	3.6	2.9
I/σ(I)	5.4	4.7	4.8	6.1	4.1	4.1	4.9	2.5

[a] Wiggler beamline of EMBL at DESY, Hamburg, equipped with a Mar CCD detector

[b] P1, P2, P3 = Peak wavelengths 1, 2 and 3

E1, E2 = Edge wavelengths 1 and 2 (point of inflection)

H1, H2 = High-energy remote wavelengths 1 and 2

L1 = Low-energy remote wavelength

[c] The inflection point and peak wavelengths were collected in inverse beam mode, whereas the remote wavelengths were collected at the low energy side of the Se edge where there is little anomalous signal and, therefore, no inverse beam data were collected

[d] R_{merge}, [e] R_{rim}, [f] R_{pim}: for details see Section 2.2.12

3.2.5.2 Structure determination by MAD phasing

***TGEV M^{pro}*:** Crystals grown from the SeMet-derivatized protein formed the basis for solving the phase problem by multiwavelength anomalous diffraction (Hendrickson *et al.*, 1990). Each of the ten sulfur atoms replaced by selenium in the SeMet protein of 302 amino-acid residues was expected to diffract anomalously at multiple wavelengths (Section 2.2.10). Crystals from the SeMet-substituted protein proved to be isomorphous to the native ones. Locating the selenium positions was first tried by taking one peak wavelength, but probably due to the low redundancy (3.8) (Table 3.5), this did not succeed. Therefore, many data sets in different combinations were merged, so that high data redundancy was achieved in order to improve the signal-to-noise ratio. By increasing the redundancy through merging of various data sets, the precision of the averaged intensities (as described by R_{pim}) was increased sufficiently. The average anomalous signal from all the measured peaks and edge wavelengths were calculated yielding more accurate diffraction data (Table 3.5).

Table 3.5 Solutions from Shake & Bake (TGEV M^{pro})

Data sets	No.of Sol/trials	R_{min} (%)	CC[a]	No.of sites	I/σ	R_{pim} (%)	$R_{anom.}$ (%)	Redund[b].
P1	0/1000	81	0.19	60	5.4	6.2	10.5	3.8
P1 + P2	6/3000	60	0.19	60	4.7	4.8	12.5	7.5
P1 + P2 + P3[c]	77/2300	52	0.47	60	9.5	4.3	14.0	11.5
H1 + H2[d]	1/4000	62	0.35	60	11.8	4.1	9.8	6.5
P1 + P2 + P3 + E1 + E2[c,e]	105/5000	49	0.51	60	13.6	3.9	12.8	18.7

[a] Correlation coefficient (The correlation coefficient is a number between 0 and 1. It is a measure of how well the predicted values from a forecast model "fit" with the real data).

[b] Average redundancy

[c] P1, P2, P3 = Peak wavelengths one, two and three

[d] H1, H2 = High energy remote wavelengths one and two

[e] E1, E2 = Edge wavelengths one and two (point of inflection)

The direct methods program SnB v2.0 (Weeks & Miller, 1999) was chosen to solve the phase problem. SnB is a multi-trial direct methods program. The program is based around the dual-space Shake & Bake algorithm, which alternates reciprocal space phase refinement with peak picking in real space. Many attempts at solving the structure were carried out. Each attempt (or trial) was started from a random set of atoms, which were then subjected to a number of cycles of phase refinement. A criterion, known as the minimal function, is then used to assess each trial and identify possible solutions. The algorithm is computer-intensive, because of the Fourier transforms that are required to continually switch between real and reciprocal space. Later each trial in *SnB* v 2.0 was processed for 120 cycles of dual-space refinement. After 1913 trials had been processed, one trial was identified that had a lower minimal function value (Guo *et al.*, 1991) (R_{min}=0.491, CC=0.51) compared to the other 1912 trials (R_{min} = 0.76-0.495, CC = 0.48). This indicated that a solution for the selenium substructure could be clearly determined. All of these initial trials were carried out using an electron-density grid size of 0.9 Å and minimum E/σ(E) = 2.55 (minimum E/σ(E) is the minimum signal-to-noise ratio for reflections to be used for phasing). Out of the 105 solutions, the first five were the top peaks from the initial solution. Although the three merged data sets collected at three peak wavelengths gave a better structure solution, the solution from five merged data sets on an average was better than the ones from three data sets because of superior quality of the data, resulting in a better R_{min}. The higher the redundancy

(or the lower R_{pim}), the higher will be the correlation coefficient and the lower the R_{min} of the correct solution (Weiss, 2001). Therefore, it is not number of trials used in *SnB,* but more importantly, the quality of data used for solutions. The positions of the best 60-atom solution from *SnB* were examined for NCS. In total, 37 positions were found to obey a 6-fold NCS. This symmetry predicted a further 11 positions (Fig. 3.9), which were not found initially from *SnB*. All 48 positions were used in MLPHARE (Otwinowski, 1991) for phasing, followed by solvent flattening and 2-fold NCS averaging in DM (Cowtan & Main, 1996), according to the standard protocols. The resulting electron density maps were of sufficient quality for chain tracing (Fig. 3.10).

Parts of the first monomer were built manually into the experimental electron density map. The other monomers were generated by NCS. The preliminary model was used as input to Arp/warp (Perrakis *et. al*., 1999) but manual intervention was still mandatory.

Therefore, initial rounds of refinement involved a combination of CNS and Arp/wArp. At some places (mostly loop and helix regions), the electron density was not good enough to build the side chains. After refining about 1000 residues out of 1800 using CNS, Arp/wArp was used for automatic building of molecules. All the bits and pieces resulting from Arp/wArp were joined together utilizing the known NCS and were put back into CNS. This was repeated for several cycles and it helped to automatically build of 200 more residues. The procedure also gave clearer electron density for the side chains. All the following refinement cycles involved only CNS. Non-crystallographic symmetry restraints were applied during the initial stages of refinement at low resolution and later gradually released as the resolution limit was extended to 1.96 Å.

All six copies (designated A - F) of the TGEV proteinase in the asymmetric unit of the crystal could be built into well-defined electron density, which covered 301 amino acid residues of each monomer, except monomer A, E, and F that lacked electron density for residue 300 as well. The final model comprises 1798 amino acid residues, 27 sulfate ions, 6 MPD molecules, 9 dioxane molecules and 1006 water molecules.

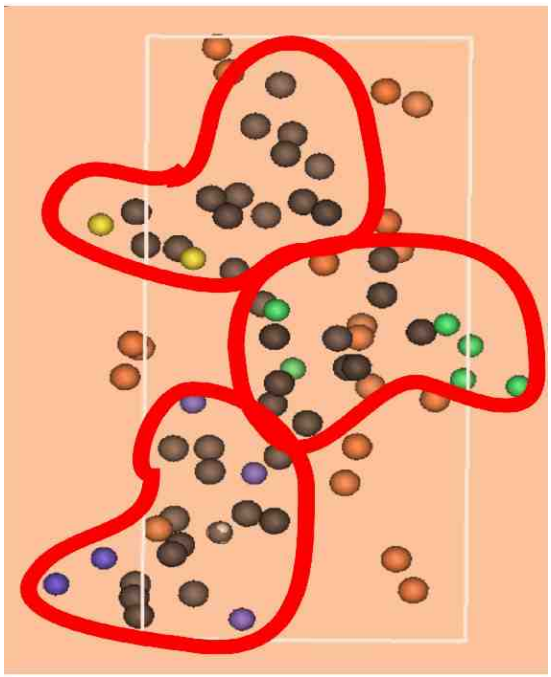


Fig. 3.9 Selenium substructure as determined by *SnB* and NCS. Brown spheres: wrong positions found by *SnB*. Dark brown spheres: correct positions found by *SnB*. Colored spheres (yellow, purple and green): additional positions determined by NCS.

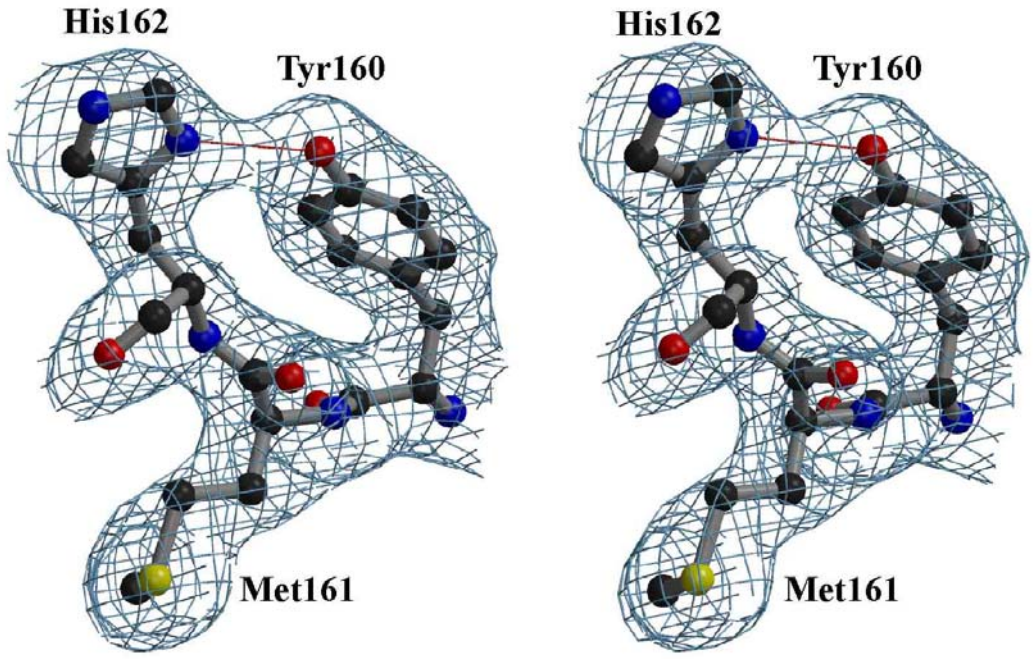


Fig. 3.10 Stereo view of a representative part of the electron density map. The ||2Fo|-|Fc|| electron density map (1.96 Å resolution, contoured at 1σ above the mean) corresponds to TGEV M^{pro} residues 160-162 (Tyr-Met-His), a conserved motif in coronavirus main proteinases. The strong H-bond (2.93 ± 0.096 Å) (average over all monomers) between the Tyr160 hydroxyl group and His 162 $N^{\delta 1}$ is indicated.

3.2.5.3 Molecular replacement method for the HCoV M^{pro} structure

60% sequence identity between TGEV M^{pro} and HCoV main proteinase made TGEV M^{pro} a good model to solve the HCoV M^{pro} structure. The structure was determined by the method of molecular replacement (Rossmann, 1990) using the AMoRe software package (Navaza, 1994). The data from the native Δ301-302 protein crystal were not good enough to solve the structure therefore, four data sets collected from different wavelengths were merged together of the SeMet-derivatized crystals of Δ301-302 HCoV M^{pro}. These data had not proved useful, as the selenium signal was weak. The four merged data sets gave highly redundant data to solve the Δ301-302 HCoV M^{pro} crystal structure by molecular replacement method (AMoRe). Since from packing consideration (Matthews, 1968), the content of one asymmetric unit of HCoV M^{pro} crystals was assumed to be two molecules, one monomer of the TGEV M^{pro} was used as a search model. For this, the TGEV monomer model was mutated to the HCoV sequence and then refined using program CNS (Brünger 1998a) to remove bad contacts. Diffraction data in the resolution range of 10 – 4 Å and 8 – 4 Å were used for rotation and translation function searches, respectively. Using a Patterson cut-off of 30 Å, a list of 10 rotation function peaks was obtained, with the top peak having a correlation coefficient of 0.116. The translation function search yielded a solution with a correlation coefficient of 0.212, fixed on the first solution from the rotation function search. After performing rigid-body refinement using the program AMoRe (Navaza, 1994) for two solutions, the correlation coefficient increased to 0.30. A summary of the molecular replacement solution for the HCoV M^{pro} structure is given in Table 3.6.

Table 3.6 Structure solution by molecular replacement: HCoV M^{pro}

Resolution range	10.0 - 4.0 Å
Rotation and translation function (1st monomer) (fractional)	
Best solution	α =21.64°, β = 59.58°, γ = 256.95° tx = 0.483, ty = 0.000, tz = 0.250 Å
Correlation coefficient	0.217
R-factor	51.9%
Rotation and translation function (2nd monomer) (fractional)	
Best solution	α = 319.92°, β = 79.38°, γ = 5.39° tx = 0.054, ty = 0.481, tz = 0.785 Å
Correlation coefficient	0.213
R-factor	52.1%
Refinement of combined solution	
Monomer 1	α = 21.80°, β = 60.40°, γ = 257.02° tx =0.478, ty = -0.002, tz = 0.250 Å
Monomer 2	α = 320.45°, β = 79.89°, γ = 5.89° tx = 0.057, ty = 0.482, tz = 0.784 Å
Correlation coefficient	0.30
R-factor	48.8%

3.2.6 Refinement and model building of the HCoV M^{pro} structure

The output model from AMoRe was subjected to rigid-body refinement using CNS (Brünger *et al.*, 1998b) and interspersed with cycles of manual inspection and rebuilding in the resolution range of 25-2.6 Å. A random set containing 4% of the total data was excluded from the refinement, and the agreement between calculated and observed structure factors corresponding to these reflections was used to monitor the course of the refinement (R_{free}, Brünger, 1992a). The model fit against the electron density, was visually checked using the program 'O' (Jones *et al*., 1991). The refinement cycles consisted of simulated annealing (Brünger *et. al*., 1990), followed by conjugate-gradient minimization of atomic coordinates and isotropic temperature factors. The first round of positional and temperature factor refinement lowered the R-factor to 36.3% and the R_{free} to 44.7%. At this stage, SigmaA-weighted maps were calculated using program SIGMAA (Read, 1986) and careful examination of the maps allowed corrections to be incorporated into the model. There was no

density for residues 212-228, 251-260, and 269-276 in each monomer in the initial maps. Therefore, these residues were removed from the model.

Alternating cycles of manual rebuilding, conventional positional refinement and simulated annealing, using the slow-cool protocol as implemented in CNS (Brünger *et al*., 1998b), allowed some of the missing residues to be placed in the density. For more accurate main-chain and side-chain tracing the crystallographic refinement was continued in a combination with OMITMAP. So, first the omitmap cycles were executed followed by refinement cycle. After each round of refinement, the 2Fo-Fc and Fo-Fc maps were calculated and the model was visually checked using 'O' (Jones *et al*., 1991). Many cycles of OMITMAP were executed, each time building an appended segment of polyalanine model and side chains into the electron density maps. The OMITMAP is calculated in order to reduce the effects of model bias. In the case of unclear density of the model, an omit map that covers the entire molecule is most useful. It is not possible to exclude the entire model (at most 10% can be omitted), instead small regions of the model are systematically excluded. A small map is made covering the omitted region. These small maps are accumulated and written out as a continuous map covering the whole molecule (or the defined region) (see Section 2.2.13).

Finally all 300 amino acid residues for each monomer in the asymmetric unit of the crystal could be built into electron density. The structure of the HCoV M^{pro} contains two copies (designated A, B) in the asymmetric unit of the crystal. There were no residues lacking electron density. The final model comprises 600 amino acid residues, 2 dioxane molecules and 221 water molecules.The refinement statistics are summarized in the Table 3.7.

3.2.7 *Quality of the model*

***TGEV M^{pro}*:** All the monomers are essentially complete, except for a few disordered residues at the carboxy terminal segment (residues 300-302 missing for monomers A, E and F, residues 301-302 in B and C, and residue 302 in D). The refinement converged to a final R factor of 0.210 and a free R factor of 0.256. The final model exhibited good stereochemical geometry. A Ramachandran plot (Fig. 3.11A) calculated with PROCHECK (Laskowski *et al*., 1993) showed that 89.1% of the residues are located in the most favored regions and 10.3% in the additionally allowed region; Asn70, Asn71, and Ser279 are in generously allowed regions (0.6%) but have good electron density, and there are no residues in the disallowed region. For evaluating the quality of structure-factor data and their agreement with the atomic

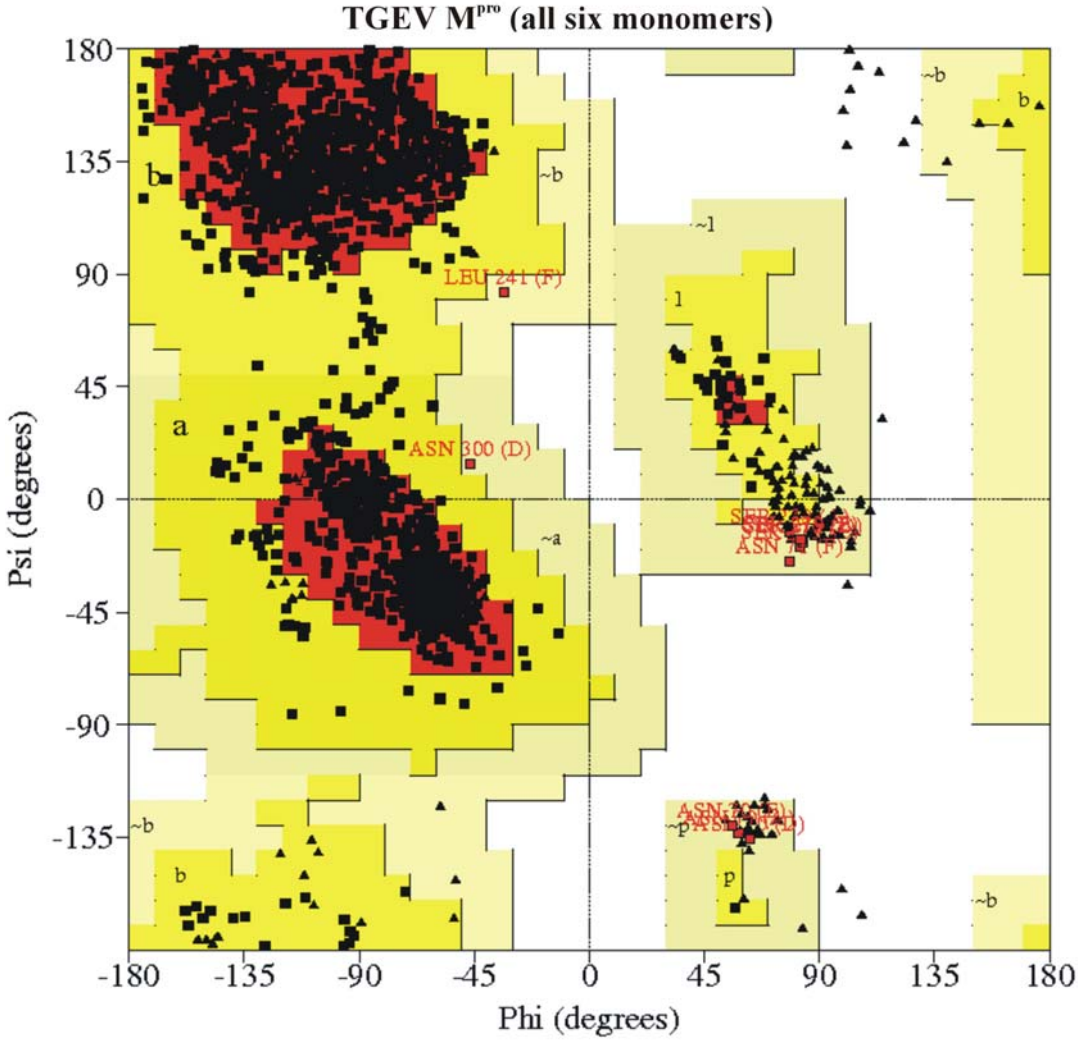


Fig.3.11A Ramachandran plot for the final model of TGEV M^{pro}. Regions are delineated as follows: red – most favored; regions a, b, p and l – additionally allowed (yellow); ~a, ~b, ~p and ~l –generously allowed (light green); white – disallowed. Glycine residues are represented as triangles; non-glycine residues as squares.

model, the program SFCHECK (Vaguine *et al*., 1999) was used, which yielded a correlation factor of 0.942 between the model generated and the experimental structure factors.

***HCoV M^pro*:** The final model consists of one dimer in each asymmetric unit of HCoV M^{pro}. There is a crystallographic symmetry between the monomers and both the monomers are essentially complete. The refinement converged to a final R factor of 0.219 and a free R factor (Brünger *et al*., 1992) of 0.283, with good stereochemistry. A Ramachandran plot (Fig. 3.11B) calculated by PROCHECK (Laskowski *et al*., 1993) showed 85.1% of the non-glycine

amino acid to be in the most favored regions of the Ramachandran plot, and 15.5% are in additionally allowed regions. The experimental structure factor data and their agreement with the calculated atomic model showed a correlation factor of 0.868, computed using the program SFCHECK (Vaguine *et al.*, 1999).

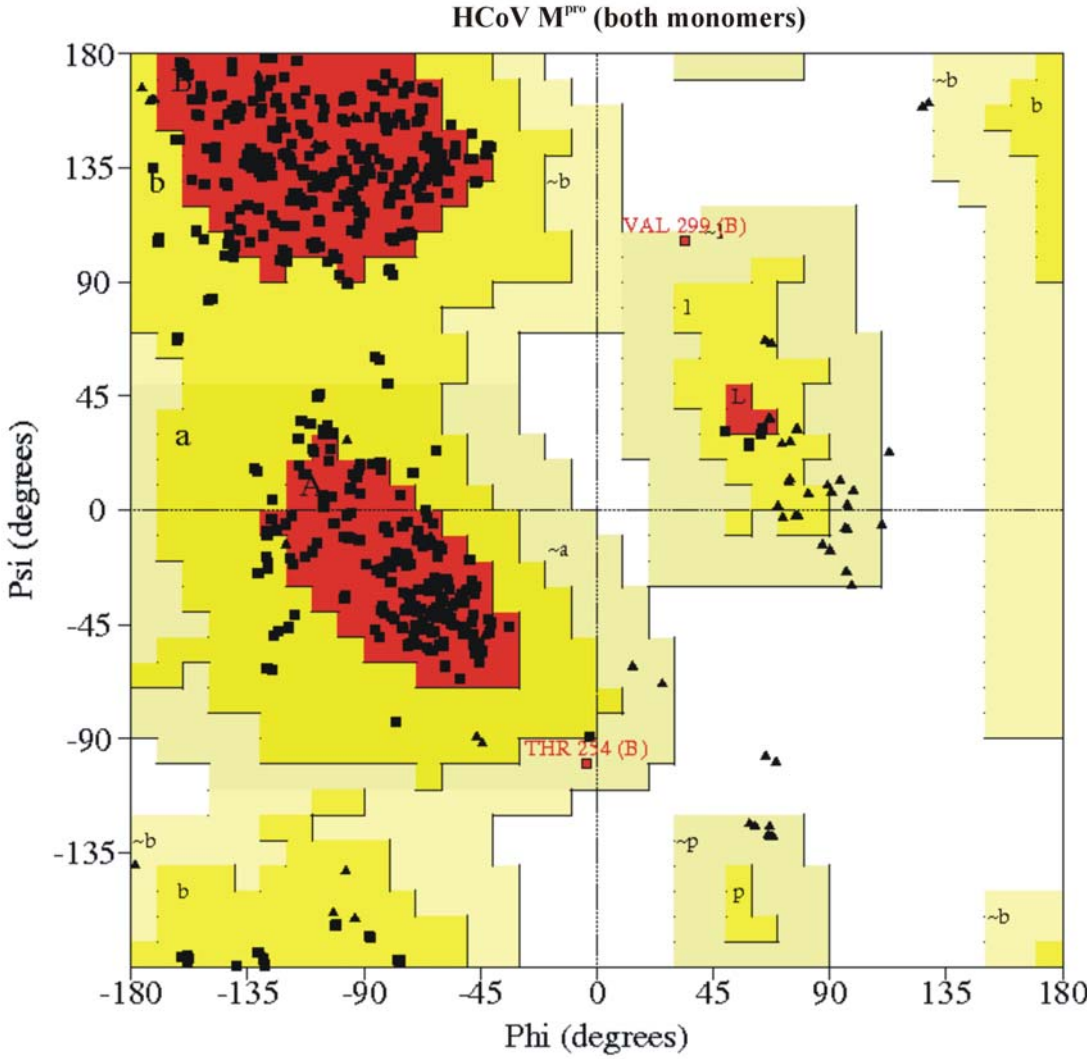


Fig. 3.11B Ramachandran plot for the final model of HCoV M^{pro}. a, b, p and l – additionally allowed (yellow); ~a, ~b, ~p and ~l –generously allowed (light green); white – disallowed. Glycine residues are represented as triangles; non-glycine residues as squares.

Table 3.7 Phasing statistics, refinement statistics and model quality

Phasing	**TGEV M^{pro}**	**HCoV M^{pro}**
FOM[a] before solvent flattening	0.48	
FOM[a] after solvent flattening (no averaging)	0.72	
FOM[a] after solvent flattening (with averaging)	0.79	
Refinement		
Resolution range (Å)	50 – 1.96	25 – 2.6
R factor[b]	0.210	0.219
R_{free}	0.256	0.283
No. of non-hydrogen atoms (average B value (Å^2))		
Protein (main chain)	7198 (46.1)	2402 (27.0)
Protein (side chain)	6613 (47.2)	2192 (27.7)
Water	1006 (50.3)	221 (24.9)
MPD	48 (67.6)	-
Sulfate	135 (57.1)	-
Dioxane	54 (71.7)	12 (58.39)
Rms deviation from ideal geometry		
Bonds (Å)	0.017	0.012
Angles (^{o})	1.9	1.5
Improper dihedral angles (^{o})	1.16	0.74

[a] FOM = figure of merit (the overall probability to determine the phase is represented by figure of merit).

[b]R-factor = Σ ($|F_o|$ - k$|F_c|$) / Σ $|F_o|$, where k is the scale factor

3.3 Structures of the M^{pro}

3.3.1 Quaternary structure

Coronavirus proteinase is found to be a monomer in solution by the gel filtration method (Ziebuhr *et al*., 1997a). However, the quaternary arrangement of TGEV and HCoV proteinases in the crystal is a homodimer. Three dimers are found in the asymmetric unit (monomers A and B, C and D, E and F) of TGEV M^{pro}, whereas there is only one dimer in the asymmetric

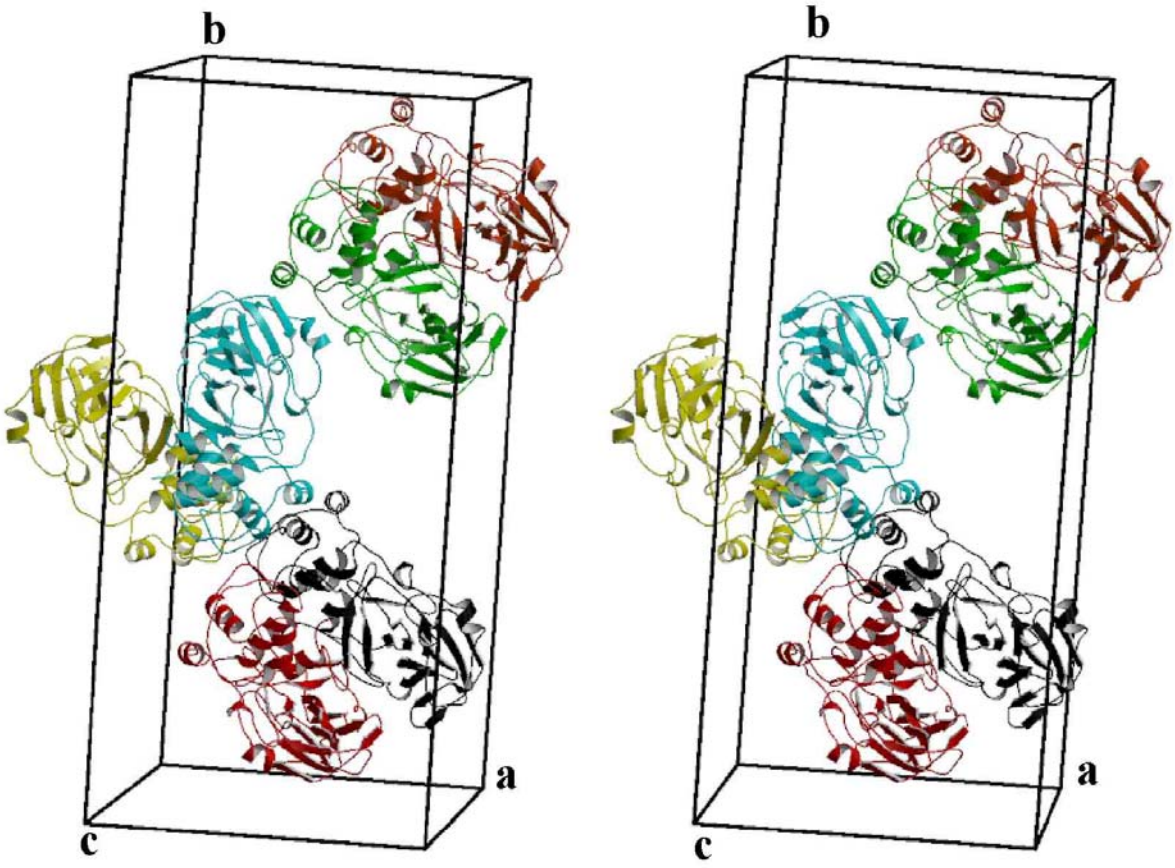


Fig. 3.12 Stereo depiction of the six molecules (three dimers) of TGEV M^{pro} in the asymmetric unit. The monomers A to F are shown in different colors; A – red, B – black, C – green, D – orange-red, E – yellow and F – cyan. Note the twofold symmetry axes between the monomers in each of the dimers, and between the two lower dimers in the figure (AB and EF). Each of the monomers measures approximately 40 Å x 22 Å x 70 Å.

unit of HCoV M^{pro} (monomer A and B). All dimers have approximate C_2 symmetry (non-crystallographic proper twofold axes) and about 1580 (±199) $Å^2$ of each monomer, *i.e.* 11 – 12% of its solvent-accessible surface, are buried upon dimerization. Figure 3.12 shows the arrangement of the dimers in the asymmetric unit of TGEV M^{pro}. The dimers AB and EF are most similar (rmsd for 598 C^{α} pairs = 0.49 Å), whereas dimer CD deviates somewhat more from dimers AB and EF (rmsd for 598 C^{α} pairs = 0.79 and 0.81 Å, respectively). The occurrence of three independent dimers in the asymmetric unit of the TGEV Mpro crystals and of one dimer in the HCoV M^{pro} structure demonstrates the relevance of the dimeric state of these proteinases. Interestingly, with chymotrypsin and picornavirus proteinases no dimer formation was observed, with the exception of poliovirus 3C proteinase (Mosimann *et al.*, 1997). In the latter case, the dimer is created by formation of a large β-sheet via β-strands e2I

of both monomers. In TGEV and HCoV proteinases, however, other regions are involved (see below) and no contiguous β-sheet is formed. Moreover, the C-terminal domain contributes to the dimer interface and might be the driving force for dimerization.

3.3.2 Tertiary structure

***TGEV M^{pro}*:** Each monomer has approximate dimensions of 40 Å x 22 Å x 70 Å, and is folded into three domains, the first two of which are antiparallel β-barrels reminiscent of those found in serine proteinases of the chymotrypsin family (Fig. 3.43). Residues 8 to 102 form domain I, and residues 103 to 182 make up domain II. The connection to the C-terminal domain III is formed by a long loop comprising residues 183 to 198. Domain III (residues 199 to 302) contains a novel arrangement of five α-helices (see below). A cleft between domains I and II, lined by hydrophobic residues, constitutes the substrate-binding site. The catalytic site is situated at the center of the cleft.

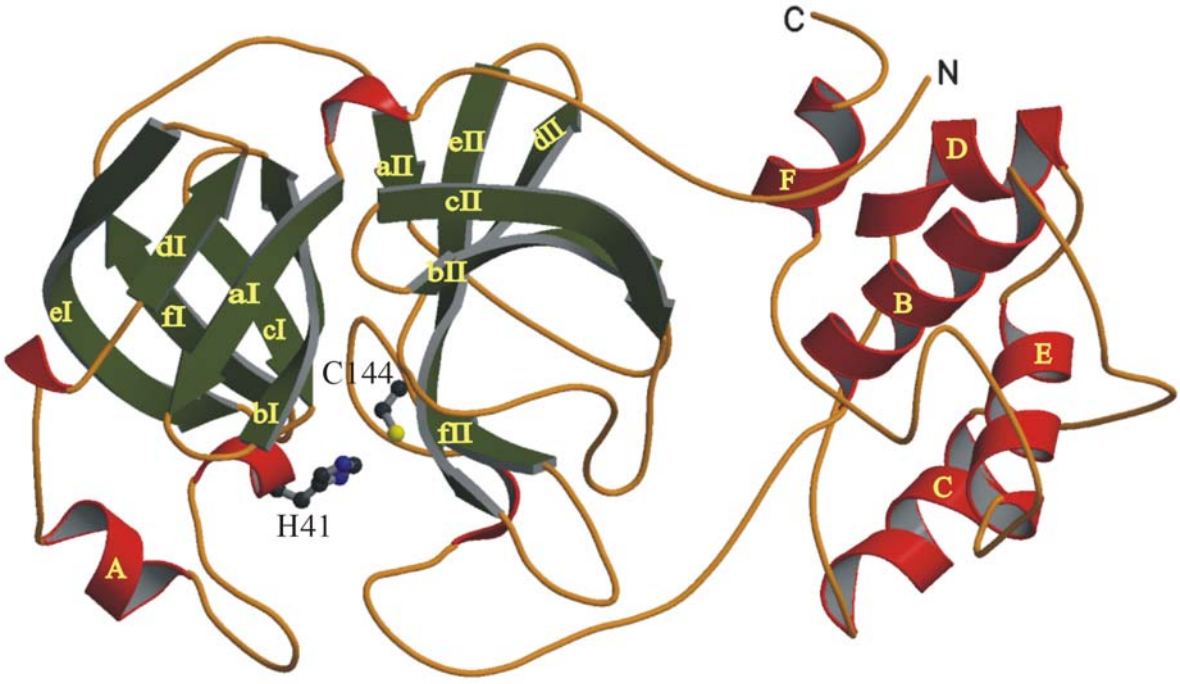


Fig. 3.13 Diagram showing the overall fold of TGEV M^{pro} with the two β-barrel domains and the α-helical C-terminal domain. The β-barrels of each domain I and II are composed of six-stranded β-sheets (green). Domain III is mainly composed of α-helices (red). Active-site residues Cys144 and His41 are depicted in a ball-and-stick mode.

The interior of the β-barrel of domain I consists entirely of hydrophobic residues. A short α helix (helix A; Tyr53 – Ser59) closes the barrel like a lid (Fig. 3.14). Domain II is smaller than domain I and also smaller than the homologous domain II of chymotrypsin and hepatitis A virus (HAV) $3C^{pro}$ (Tsukada & Blow, 1985; Bergmann *et al.*, 1997). Several secondary structure elements of HAV $3C^{pro}$ (strands bII and cII and the intervening loop) are missing in TGEV M^{pro} (see Section 3.5; Fig. 3.42 and 3.43).

Also, the domain-II barrel of the TGEV M^{pro} is far from perfect (Fig. 3.13). The segment from Gly135 to Ser146 forms a part of the barrel, even though it consists mostly of consecutive loops and turns. In contrast to domain I, a structural alignment of domain II has proven difficult. The superposition of domains I and II of the TGEV M^{pro} onto those of the HAV $3C^{pro}$ yields an rmsd of 1.85 ± 0.05 Å for 114 equivalent (out of 184 compared) C^{α} pairs, while domain II alone displays an rmsd of 3.25 ± 0.28 Å for 57 (out of 85) C^{α} pairs.

Domain III is composed of five, mostly antiparallel, α-helices and the loops connecting them. The crossover angles are ~90° between helices B and E, ~30° between B and D, ~20° between C and E, and ~80° between E and F, whereas C B and B F are parallel to each other (Fig. 3.20). Hydrophobic side chains mediate interhelical contacts.

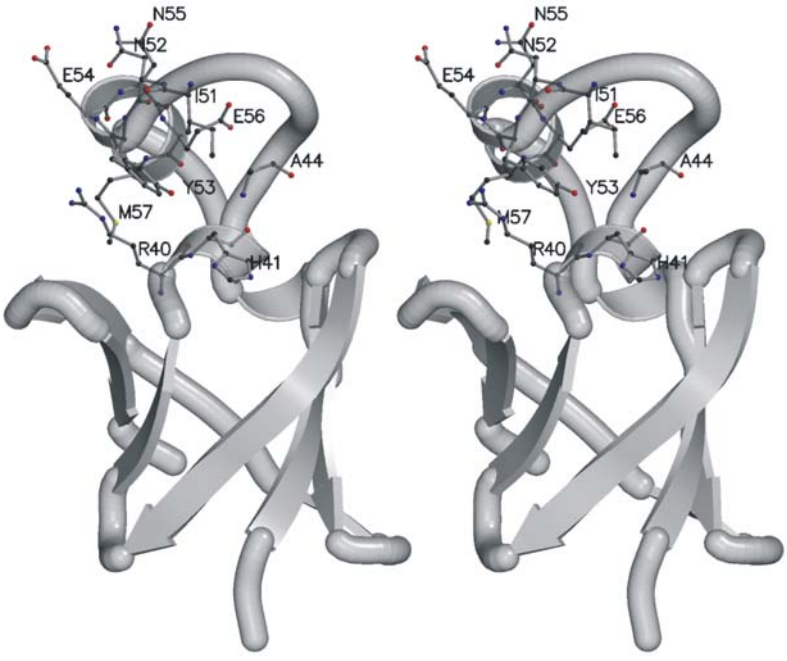


Fig. 3.14 Stereo diagram showing the hydrophobic environment of Tyr53 in TGEV M^{pro}, which forms a part of helix A that closes the domain I β-barrel like a lid.

The loops between the helices are quite long and fill up most of the interstitial space of domain III. The N-terminal segment (residues 1 –7) of the polypeptide chain folds onto domain III, placing the amino terminus of the protein within 17.0 (±2.7) Å of the C-terminus. Database searches (Holm & Sander, 1993; Gilbert *et al.*, 1999) did not reveal other proteins or protein domains with the same topology as TGEV M^{pro} domain III (Section 2.2.15).

The C-terminal domain of the TGEV M^{pro} has been shown to be, directly or indirectly, involved in the proteolysis reaction, since deletion of the entire domain leads to a nearly complete loss of proteolytic activity in a peptide cleavage assay (Anand *et al*., 2002a). The formation of the TGEV M^{pro} dimer involves, among others, interactions between domains III of the monomers across the non-crystallographic twofold axis. This interface is anchored by only two hydrogen bonds, between the amide group of Gly281 (molecule A) and the main-chain oxygen of Ser279 (molecule B), as well as its symmetry mate, Gly281B...Ser279A (3.22 ± 0.37 Å, averaged over all six monomers). Involving an area of only 337 ± 45 $Å^2$, the domain III – domain III interface appears to be the consequence rather than the cause of other intermolecular interactions described below (Section 3.4.7), in particular between domains II and III of one monomer and the N-terminal residues of the other.

***HCoV M^{pro}*:** The HCoV M^{pro} structure has a fold similar to that of TGEV M^{pro}; it is divided into three domains as well. The monomer measures ~36 Å x 26 Å x 69 Å. The two monomers (designated A and B) in the asymmetric unit have the same mutual arrangement as in TGEV M^{pro}. The monomers are oriented at an angle of 90^o to each other (Fig. 3.15). Domain I consists of residues 8 to 99, and domain II of residues 100 to 183. They comprise six β strands each, folded into β-barrels as in chymotrypsin-like proteases. As in TGEV M^{pro}, these two domains provide the residues that define the specificity for the substrate that binds in a cleft between them. There is a long loop from residue 184 to 199 connecting domain II to the C-terminal domain. The C-terminal domain is a globular cluster of five helices. Domain III of monomer B is rather flexible, with a higher overall average B factor compared to domain III of monomer A (see also Fig. 3.26). The C-terminal domain has been implicated in the activity of the protease. Deletion of 81 residues from the C-terminus of HCoV M^{pro} has been reported to lead to total loss of proteolytic activity (Ziebuhr *et al*., 1997b) in the standard peptide cleavage assay. Similar but more detailed results have been obtained for TGEV M^{pro} in more

sophisticated experiments designed on the basis of the crystal structure described in this work (Anand *et al*., 2002a; see below).

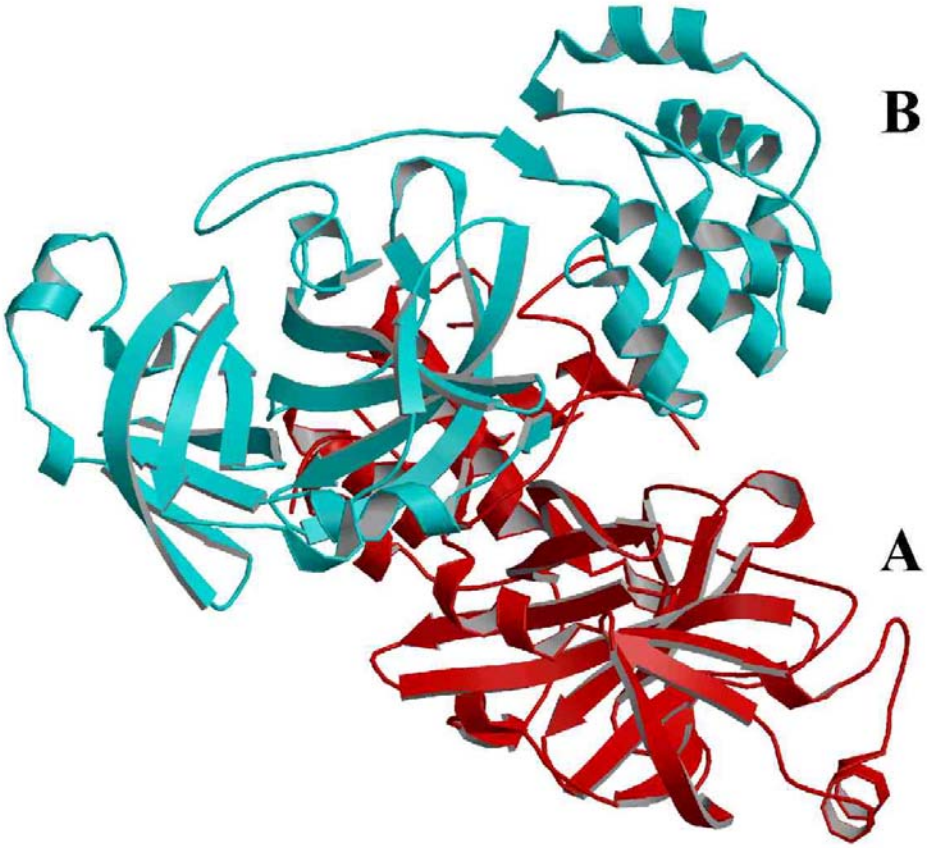


Fig. 3.15 Arrangement of monomers in the asymmetric unit of HCoV M^{pro}. Both monomers are lying perpendicular to each other (~90^{o}). A similar arrangement of monomers is seen in the dimer of TGEV M^{pro}.

3.3.2.1 Comparison between the monomers

***TGEV M^{pro}*:** Figure 3.16 shows the C^{α} backbones of all six molecules superposed on each other using monomer A as a reference. As can be seen from the figure, the backbone trace is very similar. A superimposition of any two monomers in the asymmetric unit gives rmsd values from 0.28 to 0.43 Å on the basis of 299 C^{α} pairs. The molecules superimpose very closely near the active site and larger deviations occur for residues that are radially further out. The surface near the active site, presumably involved in molecular recognition, is retained in all the molecules in spite of non-identical environments in the crystal.

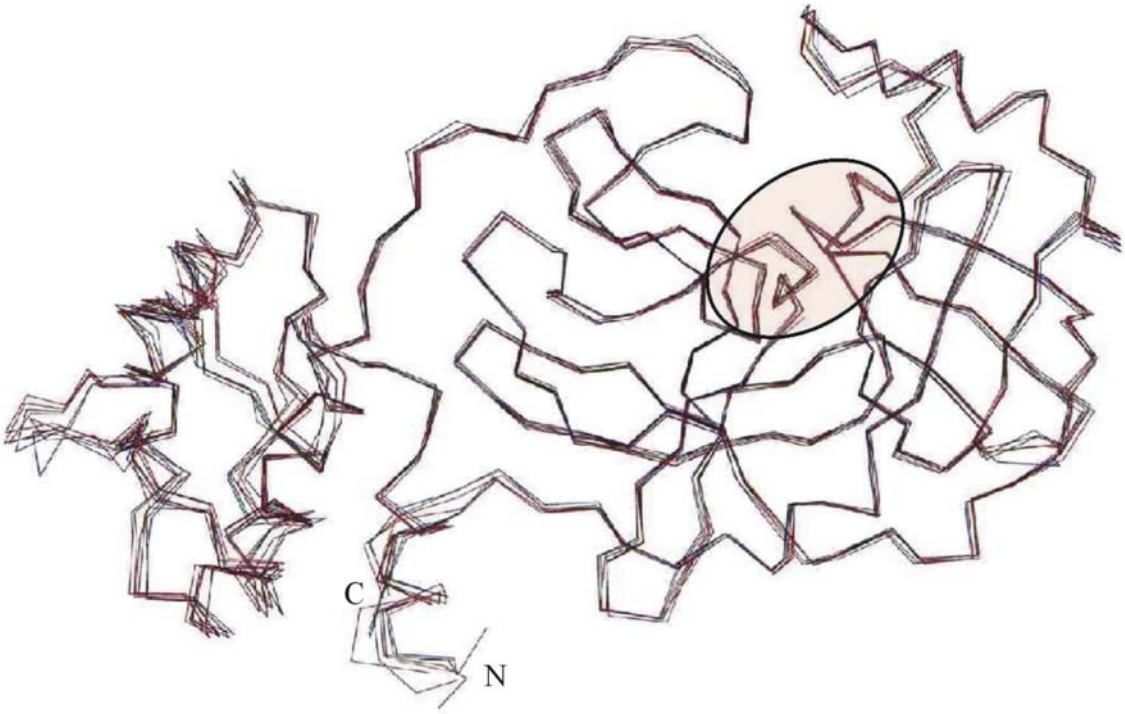


Fig. 3.16 Monomers B to F of TGEV M^{pro} superimposed on monomer A. The core part has an rmsd of 0.29 (± 0.09) Å for 130 equivalent C^{α} atoms. The largest deviation occurs at the N-terminal segment (residues 1-4) and at the C-terminal segment (residues 294-300). The active-site region is approximately indicated by a light orange circle. N- and C-terminals are labeled.

The core of domains I and II (defined as the region around the active-site cleft between the two domains) displays an rmsd of 0.29 (± 0.09) Å for 130 equivalent C^{α} atoms. If all 299 well-determined C^{α} positions are included, the average rmsd for all monomers is 0.57 (± 0.18) Å (monomer A as reference). Among all monomers, the pairs A-E and B-F are most similar to one another (rmsd 0.24 Å and 0.28 Å, respectively), while B-C displays the largest difference (rmsd 0.43 Å). The difference in the rmsd values indicates that there is some variation in the loop regions, which is probably due to crystal packing effects. The largest deviations of the main-chain trace are in: i) the N-terminal segment from residues 1 to 4 (average rmsd: 1.69 ± 0.91 Å), ii) the flexible surface loop from residues 216 to 225 (average rmsd: 0.99 ± 0.51 Å); iii) the C-terminus of helix E and the loop region between residues 267 and 276 (average rmsd: 0.99 ± 0.42 Å); and iv) the segment 294-300 following the C-terminal F helix (average rmsd: 1.55 ± 0.44 Å). In addition to being flexible and at the surface of the molecules, segments ii) and iii) are involved in inter-dimer crystal contacts in some but not all of the six

protomers. Surprisingly, the regions with the highest rmsd are not the regions with the highest temperature factors except for the C-terminal domain of monomer F that does have high temperature factors (~70 $Å^2$, whole model 47 $Å^2$, including all 1006 water molecules; also see Fig. 3.25). The inter-dimer interface (Fig 3.12) between monomers A and C involves mainly loops, turns and 3_{10} helices, whereas it contains mostly loop regions for the mutual packing of monomers B and E. Between monomers D and B, it is mostly helices that are involved in the interface (see Section 3.3.3.1). There are also limited contacts between monomers A and B, on the one hand, with monomers E and F, on the other. Similarly, monomer C (domain III) also has contacts with monomer F (domain I).

A domain-wise comparison of monomers of the TGEV M^{pro} reveals that domain II has the best superimposition (~0.39 Å) followed by domain I (0.48 Å) and domain III (0.79 Å). The helical domain III is quite flexible. It has many long loop regions and high thermal factors in both proteinases. Comparison along the secondary structure elements reveals the loop regions of the domain to be deviating most.

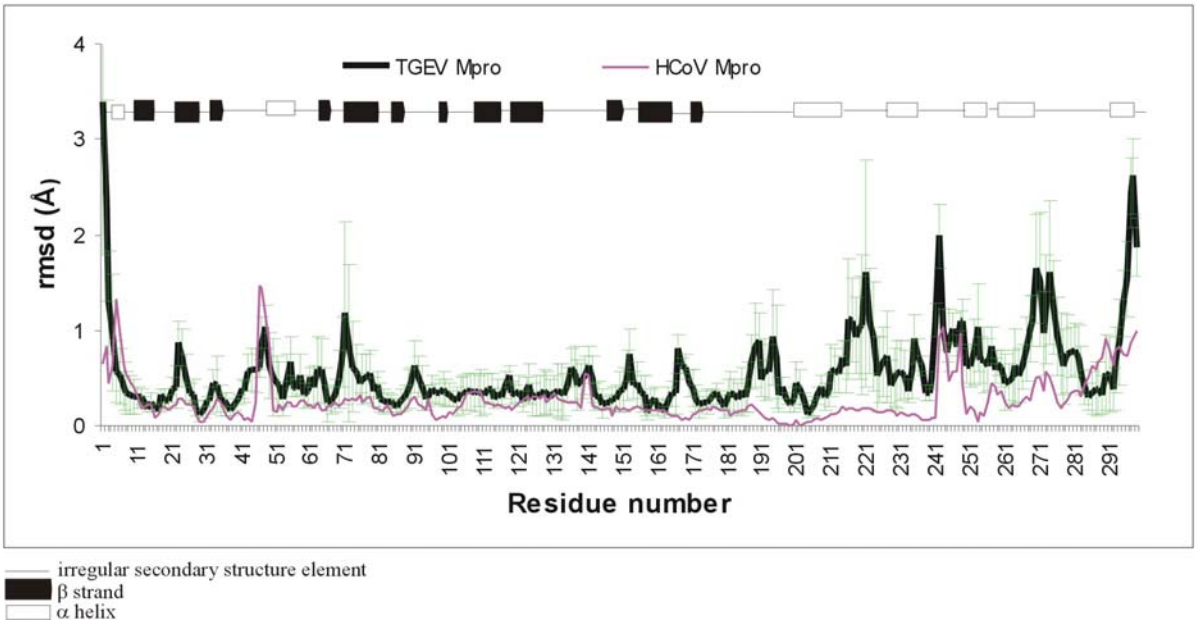


Fig. 3.17 Average rms deviation between TGEV M^{pro} monomer A and the other monomers (B-F) (black), and HCoV M^{pro} monomers A and B (monomer B on A, magenta). The standard deviation for TGEV monomers (B-F) superimposed on molecule A is shown in green. The first helix indicated in the secondary structure legend is absent in TGEV M^{pro}.

***HCoV M^{pro}*:** In the HCoV proteinase dimer, the rmsd deviation between monomer A and B is 0.28 Å for 300 pairs of C^{α} atoms. The two monomers are related by two-fold non-crystallographic symmetry. The arrangement of the monomers in the dimeric HCoV M^{pro} is similar to that seen in TGEV M^{pro}. The most different regions are the N- and C-terminal segments, the loop region 46–49 and the loop-helix region 242–256. These are also the regions of comparatively high temperature factors. The domain-wise rms deviation between the monomers are 0.27 Å, 0.19 Å, and 0.40 Å for the domains I, II and III, respectively. The average solvent-accessible area for the HCoV M^{pro} monomers is 14231 (±64) $Å^2$ and around 9% of this surface is buried on dimer formation.

3.3.2.2 Comparison between TGEV M^{pro} dimers

The general arrangement of the monomers is such that the AB pair, the CD pair and the EF pair share maximum contacts. Molecule A residues from β-strand cII, N-terminal residues, and loop regions of domain III pack against the equivalent regions of molecule B. Similar arrangements are present in all the remaining dimers. At least four inter-subunit salt bridges stabilize each of the TGEV M^{pro} dimers. Among these, two link domain III of one monomer to the N-terminus of the other: Glu286A $O^{\varepsilon 2}$-Arg4B $N^{\eta 2}$ and its symmetry mate, Glu286B $O^{\chi 2}$-Arg4A $N^{\eta 2}$ (average distance between the $O^{\varepsilon 1}$ and NH1 atoms: 5.31 ± 1.02 Å). The third one is N-terminal Ser1A N-Glu165B $O^{\varepsilon 2}$ and its symmetry mate, Ser1B N-Glu165A $O^{\varepsilon 2}$ (average distance between N and OE2 atoms: 4.72 ± 2.17 Å) (see appendix, Table 6.2.1 - 6.4). The protein has a tightly packed structure devoid of many water molecules in the interior (only 78 out 1006 for all three dimers) (see appendix, Table 6.3A, B). The average accessible-surface area of the isolated monomers is 14136 (± 112) $Å^2$. There is an average burial of 11% upon formation of the dimer(s).

3.3.3 Structural relationship between the TGEV and HCoV M^{pro}

There is a high sequence identity (60%) between the two coronavirus proteinases. Both proteins crystallize as dimers and the crystals belong to space group $P2_1$. The difference is in the number of dimers in the asymmetric unit, TGEV M^{pro} contains three dimers, whereas HCoV M^{pro} has only one. The fold of the polypeptide chain is similar in both cases. When 300 C^{α} atoms of HCoV M^{pro} monomer A, including all those which are involved in regular secondary structure elements, are superimposed onto monomer A of TGEV M^{pro}, a high overall rmsd value of 1.5 Å is obtained (Fig. 3.18). This high value is due to the most

deviating region, *i.e*. domain III (~2.55 Å). Domains I and II fit much better, with rmsd values of 1.05 Å and 0.90 Å, respectively. When domain III is excluded from the alignment, the rmsd value improves to 0.68 Å for 162 target pairs (out of 184 residues). The isolated domain III structures of the two M^{pro}s superimpose rather well (0.79 Å for 94 target C^{α} pairs), indicating that the relative orientation of this domain differs in the HCoV and TGEV M^{pro} structures.

3.3.3.1 Interface

The interface between the respective monomers of both TGEV M^{pro} and HCoV M^{pro} dimers is rather large (~2500Å^2). All three domains are involved in the monomer-monomer interface. The major portion of the interface area is however contributed by domain II, the long loop 184-199 and some smaller loops, the β-strand cII as well as two residues, Glu165 and His171, from β-strands eII and fII, respectively. Also, the N-terminal segment 1-14 from domain I and some residues from the C-terminal segment (helix F, 281 - 300) participate in the interface. Mostly main-chain amides and carbonyl oxygen atoms are making hydrogen bonds in the interface in both TGEV M^{pro} and HCoV M^{pro}.

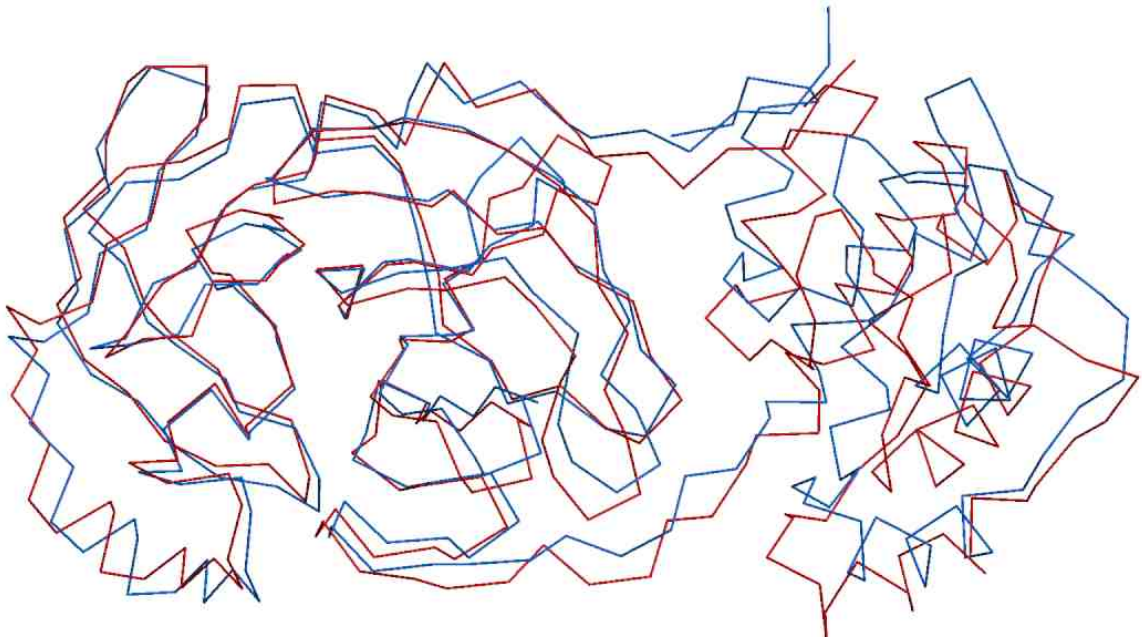


Fig. 3.18 Three-dimensional superposition of the C^{α} traces of TGEV (red) and HCoV M^{pro} (blue).

Domain III is the most divergent region between the HCoV and TGEV M^{pro}. Most of the mutations between the M^{pro}s, are also located in domain III. Domain II has the most identical residues in the structures. There is a significant difference in the number of charged

residues. Glu is in large excess in TGEV M^{pro}, as a result of which the theoretical estimate of charge is double that of HCoV M^{pro}. A look at figure 3.19 suggests that most amino acid exchanges are contiguous in three-dimensional space.

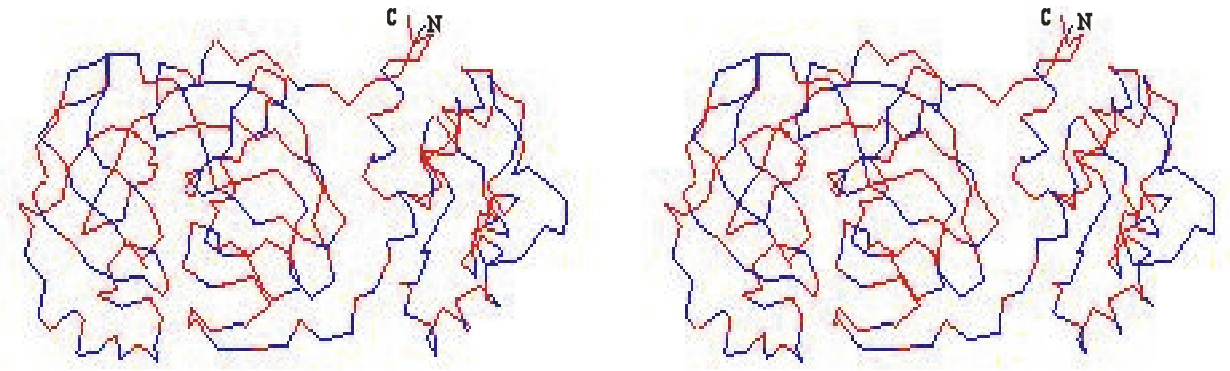


Fig.3.19 Three-dimensional locations in the polypeptide chain (red) where amino acids are exchanged (in blue) in TGEV M^{pro} to yield the HCoV M^{pro} structure. N- and C-termini are also shown.

3.3.4 Secondary structure: TGEV and HCoV M^{pro}

3.3.4.1 α-Helices

TGEV and HCoV M^{pro}s contain six α-helices and 16 short 3_{10} helices each. HCoV M^{pro} has an extra short helix at the N-terminus. The α-helices are stabilized by recurring hydrogen bonds between NH and CO groups at positions *i+4* and *i* in the sequence. The α-helices of TGEV M^{pro} are labeled as A from residue 53 to 58 in domain I, B from residue 200 to 212, C from residue 226 to 235, D from residue 248 to 254, E from residue 258 to 266 and F from residue 289 to 295 (Fig.3.20). The α-helices in HCoV M^{pro} are A' from residue 11-14 and A from residue 53 to 58 in domain I, B from residue 200 to 212, C from residue 226 to 234, D from residue 249 to 252, E from residue 258 to 268 and F from residue 289 to 297. The lengths of the helices are identical in the two structures except for helix C, D and F. In HCoV M^{pro}, the C helix is shorter by one residue at the C-terminus. Helix D in HCoV M^{pro} is shorter than in TGEV M^{pro} by three residues, one at the N-terminus and two at the other end; helix F is longer by two at the C-terminal end. The extra N-terminal helix A' in HCoV M^{pro} is replaced by a stretch of turn and 3_{10}-helical residues in TGEV M^{pro}.

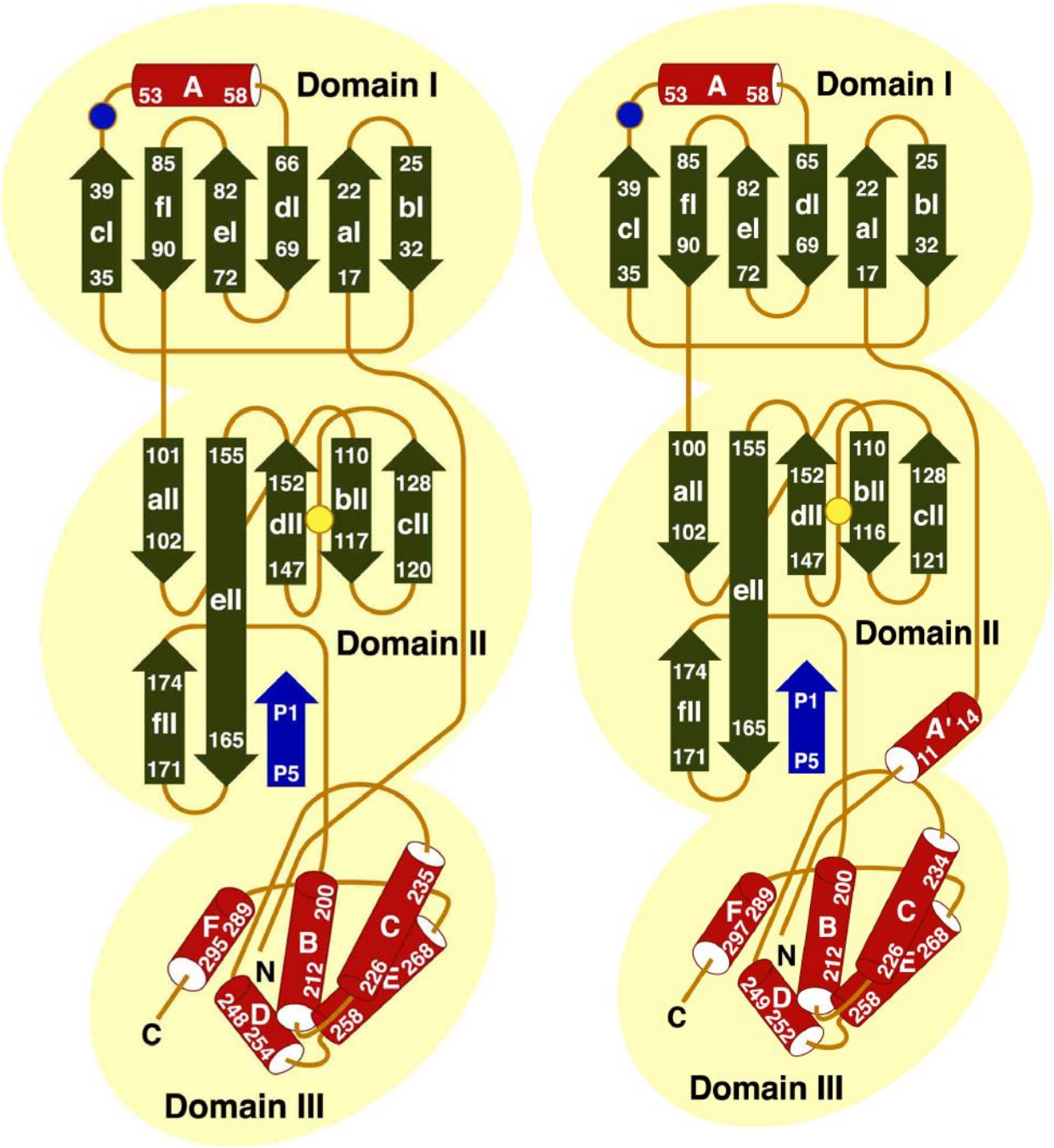


Fig. 3.20 Topological representation of the secondary structure elements of a TGEV M^{pro} (left) and HCoV M^{pro} (right) monomer. α-helices and β-strands are represented as cylinders and arrows, respectively. The N- and C-terminal residue positions of the secondary structure elements are indicated. Strands bI and cI are adjacent. The active-site residues Cys144 (yellow) and His41 (blue) are shown by circles. The positions of the N- and C-termini of the M^{pro} are indicated. Also, the presumed localization of the P5-P1 region of a model substrate is shown (blue) (for details see Section 3.4.4 and Fig. 3.31).

3.3.4.2 β-Sheets

About 1/4th of the polypeptide sequence of TGEV and HCoV M^{pro} is made up of almost fully extended strands, with ϕ,ψ angles within the wide, shallow energy minimum in the upper left quadrant of the Ramachandran plot (Fig. 3.11A). There are twelve strands in TGEV M^{pro}, six each in domains I and II. Strand aI in domain I from residue 17-22, bI from 25 to 32, cI from 35 to 39, dI from 66 to 69, eI comprises residue from 72 to 82 and fI from 85 to 90. The six β-strands in domain II are aII from residue 101 to 102, bII from 110 to 117, cII from 120 to 128, dII contains 147 to 152, eII consists of 155 to 165, fII contains 171 to 174 (Fig. 3.20). In HCoV M^{pro}, the situation is identical except for strands dI, aII, bII and cII. In dI and aII, the β-strands are longer by one residue at the N-terminus. The β-strands bII and cII are shorter by one residue at C- and N- terminus, respectively. All the strands are antiparallel except the shortest strand aII, which is parallel to the longest strand eII. The antiparallel sheets have one side exposed to solvent and the other side buried. They show an alternation of side-chain hydrophobicity in the amino acid sequence. There is a tendency toward greater hydrophobicity for the central than for the edge strands of the sheet (Sternberg & Thornton, 1977). In the TGEV and HCoV M^{pro} structures, the large antiparallel sheets roll up either partially or completely around to join the edges into a cylinder or 'barrel'. The barrels in domains I and II of TGEV and HCoV M^{pro} measure ~24 Å in length, ~15 Å in width, and ~17.3 Å in length, ~12.7 Å in width, respectively. The barrels are not uniformly cylindrical, especially that of domain II. The diameter of the barrel is dependent on the number of constituting strands and their twist – the twist being one of the most conspicuous features of the β sheet (Chothia, 1973). Residues 147 to 152 (dII) form an antiparallel sheet with strand eII; however, the strand beyond residue 159 in eII is twisted and is almost perpendicular to the adjacent dII strand. Due to this, there is no hydrogen bond from residue 159 to the neighboring dII strand. The twist can be measured by the angle at which strands on opposite sides of the barrel cross one another; that angle averages 40^{o} for 7- and 8-stranded, 30^{o} for 9-13-stranded barrels and 95^{o} for 6-stranded antiparallel barrels (Chothia, 1973), as in the case of TGEV and HCoV M^{pro}s, Strands cI, dI, eI and fI form a Greek key motif. The nomenclature of the β-strands used here is based on picornavirus 3C proteinase structures.

3.3.4.3 Reverse turns

Reverse turns are a common feature of protein structure. These conformational elements enable proteins to adopt their globular structure by significantly changing the polypeptide chain direction. Reverse turns are of many types, among which β-turns are most frequent and are composed of four residues forming a bend involving a hydrogen bond between the CO(*i*) and NH(*i+3*) group. The presence of the *i*←*i+3* hydrogen bond constrains the backbone torsion angles (ϕ,ψ) of residues in the *i+1* and *i+2* position to a limited number of combinations (Venkatachalam, 1968), giving rise to turn types of different denominations ─ I, I', II, II', III, III', etc. In some cases, however, in spite of proper (ϕ,ψ) angle combinations at *i+1, i+2* positions, the hydrogen bond is not always in a proper geometry, and a $C^{\alpha}(i)...C^{\alpha}(i+3)$ distance of less than 7.0 Å is also considered a limit for β turns in such a case (Richardson, 1981).

There are numerous β-turns in TGEV M^{pro}, mostly of the type I, I′, II and II′. Type I' and II' are mirror images of I and II, respectively. Normally, type I and II turns are far more frequent in proteins than I' and II' (Creighton, 1993). But the situation is different in TGEV M^{pro}. The conformational details are listed in Table 3.8. In turns, the side chains are usually solvent-accessible and are often hydrophilic. Additionally, they can form side chain-to-main chain hydrogen bonds, which presumably stabilize the turn. However, the β-turns rarely contain side chains that locally stabilize the turn at all four positions. In fact, many side-chains are involved in long-range interactions that help stabilize the tertiary structure. This is broadly found in TGEV/HCoV M^{pro} with some exceptions. A particularly interesting example is a type I turn that is rather hydrophobic (R216-W217-F218-V219), and present adjacent to the helix-capping Schellman motif (Schellman, 1980) in position 210-215 in TGEV M^{pro}. Although the phenyl group of F218 is buried inside the core, the Trp group is exposed to the surface and is partly covered by the side chain of Arg 275. The remaining part of the indole ring is involved in an intersubunit contact stabilizing the subunit association.

Almost half of the β turns in TGEV/HCoV M^{pro} are involved in reversing the direction of the polypeptide so that two adjacent stretches can form antiparallel strands. Strand pairs aI,bI; bI,cI; dI,eI; eI,fI; bII,cII; dII,eII; eII,fII form a strand-turn-strand segment. Of these eII, fII are connected by type I, bII,cII; and dII,eII by type I' and the rest by type II' turns.

Table 3.8 Torsion angles and distances of β-turns of monomer A in TGEV M^{pro}

Residue	CO(*i*)... NH(*i+3*) (Å)	C^{α}(*i*)... $^{\alpha}$(*i+3*) (Å)	ϕ_{i+1} (°)	ψ_{i+1} (°)	ϕ_{i+2} (°)	ψ_{i+2} (°)	Turn type
D46-T47-T48-R49	3.49	6.08	-86.16	5.83	-97.00	-5.97	I
N94-P95-N96-T97	3.32	5.36	-71.38	-11.16	-105.40	23.66	I
R130-S131-Q132-G133	2.79	5.36	-52.44	-35.86	-83.76	3.26	I
L166-G167-N168-G169	2.65	5.42	-49.10	-34.91	-80.18	11.36	I
N175-F176-E177-G178	3.12	5.54	-85.66	7.32	-93.58	-5.05	I
R216-W217-F218-V219	2.70	5.52	-40.83	-31.38	-83.38	-6.67	I
L268-N269-K270-G271	2.68	5.26	-44.48	-34.33	-80.36	-0.69	I
Y117-E118-G119-C120	2.78	5.13	34.37	57.60	93.75	-9.16	I'
E152-N153-G154-I155	2.81	5.21	51.28	43.27	80.34	-6.46	I'
F272-G273-G274-R275	3.15	5.22	61.83	31.75	92.14	-9.26	I'
I277-L278-S279-Y280	2.92	5.24	44.97	45.75	78.45	-13.01	I'
K106-A107-G108-E109	2.92	5.54	-52.37	142.30	80.39	-6.04	II
I140-A141-G142-T143	3.07	5.94	-46.36	132.74	103.97	-13.00	II
T143-C144-G145-S146	2.95	5.64	-57.28	145.51	88.98	-10.57	II
Y22-G23-N24-N25	3.08	5.45	71.47	-133.74	-95.85	4.73	II'
L32-G33-D34-E35	2.96	5.50	58.93	-132.25	-106.16	26.14	II'
K69-N70-N71-V72	2.84	5.26	58.48	-133.24	-105.36	13.02	II'
K82-G83-V84-N85	3.05	5.18	67.15	-134.05	-87.50	3.98	II'
S10-G11-L12-V13	2.95	5.41	-50.85	-37.60	-71.45	-31.25	III

The presence of a large number of strand-turn-strand segments is also a requirement for the formation of the β-barrel topology. Sibanda and Thornton (1985) have observed that type I' is predominant for antiparallel hairpin bends, which is not the case in TGEV M^{pro}. Gly is present in all these turns and its preference at the *i+1* and *i+2* positions (type I'/II') is consistent with the presence of a positive ϕ angle in one (or both) of these positions. Only a single case of type II turn is present, where the two-residue β strand (aII) is in contact with the adjacent parallel strand eII.

The presence of the all-α-helical domain III requires the presence of helix-loop-helix segments to ensure proper packing of the helices. However, in TGEV M^{pro}, there are no short loops that forms helix-turn-helix segments. In fact, all the turns that have not been discussed so far are part of a much longer loop connecting two regular secondary structural elements. A majority of these are type I turns, consistent with the previous observation of their high

frequency in globular proteins (Creighton, 1993). Interestingly, however, the two adjacent turns (residues 140-143, 143-146) that involve the active site residue Cys144, are of type II category. They have Gly at the *i+2* position, as observed earlier (Wilmot & Thornton, 1988). There are no *cis* peptide bonds in both TGEV M^{pro} and HCoV allowing for any possibility of type VI β turn formation.

All turns observed in TGEV M^{pro} are observed in HCoV M^{pro}, with four exceptions. The turn segment from residue 10 to 13 in TGEV M^{pro} is replaced by an α-helix, and the immediate neighbouring 3_{10}-helical segment is replaced by a type I turn (middle residues 15 and 16 (ϕ,ψ; -65.26, -11.95 and -97.74, -10.39 respectively) connected to strand aI in HCoV M^{pro}. The residue segment 117-120 (type I' turn) and 210 to 219 (consisting of a Schellman motif followed by a type I turn) of TGEV M^{pro} are found as simple non-hydrogen-bonded bends in HCoV M^{pro}. Residues 253 - 255 are in the type III conformation (ϕ, ψ; -52.21, -32.88; -72.94, -53.08; -50.65, -31.43) at the C-terminus of helix D in HCoV M^{pro}; the corresponding region in TGEV M^{pro} has an α-helical structure. Two examples of β turn are shown in Fig 3.21.

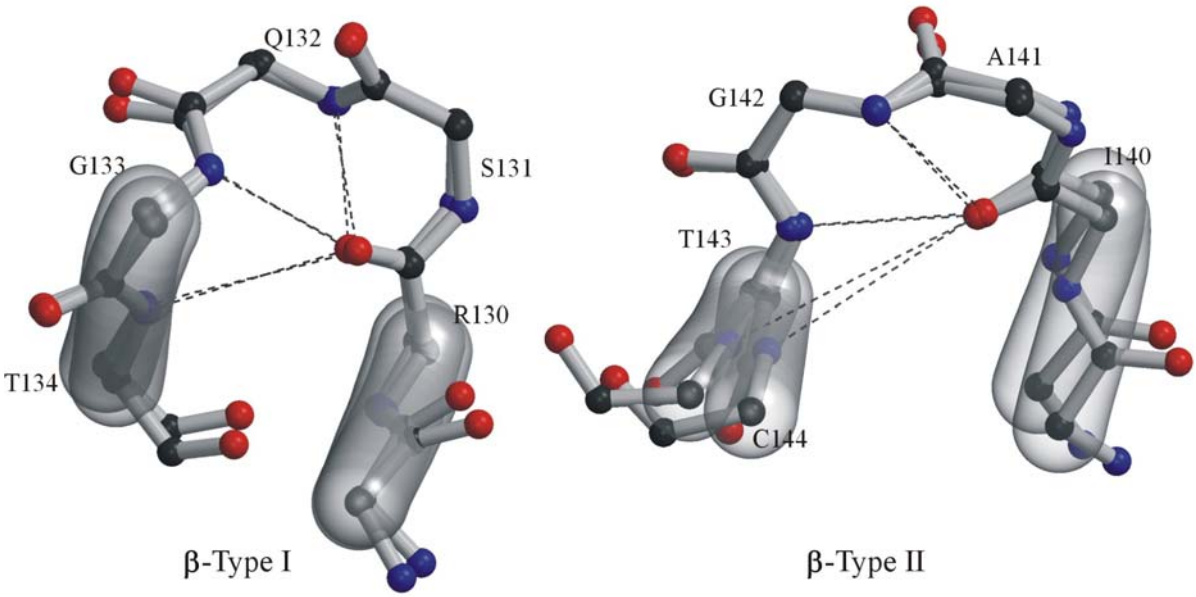


Fig. 3.21 Example for type I (left) and type II (right) β-turns in the TGEV M^{pro} and HCoV M^{pro} crystal structures superimposed on each other. Side chains are omitted for clarity.

3.3.4.4 α–Turns

Five consecutive residues in the two sequence segments A233-K234-T235-N236-S237-F238 and A251-A252-K253-T254-G255-Q256A make an intra-turn hydrogen bond between the backbone O_i and backbone NH_{i+4} and their C^{α}_{i}–C^{α}_{i+4} distances are all less than 7 Å. The former is at the C-terminus of the C-helix and the latter is connecting the D- and E-helices (Fig. 3.22). Specific backbone torsion-angle combinations are possible for α turns, which have been previously classified into I-α_{RS}, I-α_{LS}, II-α_{RS}, I-α_{RU}, I-α_{LU}, II-α_{RU}, II-α_{LU}, and I-α_{C} based on particular combinations of backbone dihedral angles (Nataraj *et al*., 1995; Pavone *et al*., 1996; Chou, 1997). Using the torsion angle criteria, the above two examples can be classified as I-α_{RS}, although the backbone angles of N236 in the first example have highly deviant torsional values. It appears that the loop segment beyond F238 is responsible for such discrepancy. Both the turns have a Lys each, indicating their highly hydrophilic nature. Gly in the helix termini as in G255 is known to form α-turn Gly motifs (Aurora *et al*., 1994). In HCoV M^{pro}, both the α-turns are also detected. Below is the table showing the residues involved and their conformation.

Table 3.9 α-Turn examples from TGEV M^{pro} subunit A

Residue	CO(*i*)... NH(*i+4*) (Å)	$C^{\alpha}(i)..C^{\alpha}(i+4)$ (Å)	ϕ_{i+1} (°)	ψ_{i+1} (°)	ϕ_{i+2} (°)	ψ_{i+2} (°)	ϕ_{i+3} (°)	ψ_{i+3} (°)	ϕ_{i+4} (°)	ψ_{i+4} (°)
A233-K234-T235-N236-S237-F238	3.78	6.46	-54.31	-28.82	-108.51	10.77	-130.26	12.78	70.7	40.40
A251-A252-K253-T254-G255-Q256	2.93	5.48	-58.86	-49.21	-66.11	-36.19	-90.05	-20-75	84.59	-5.82

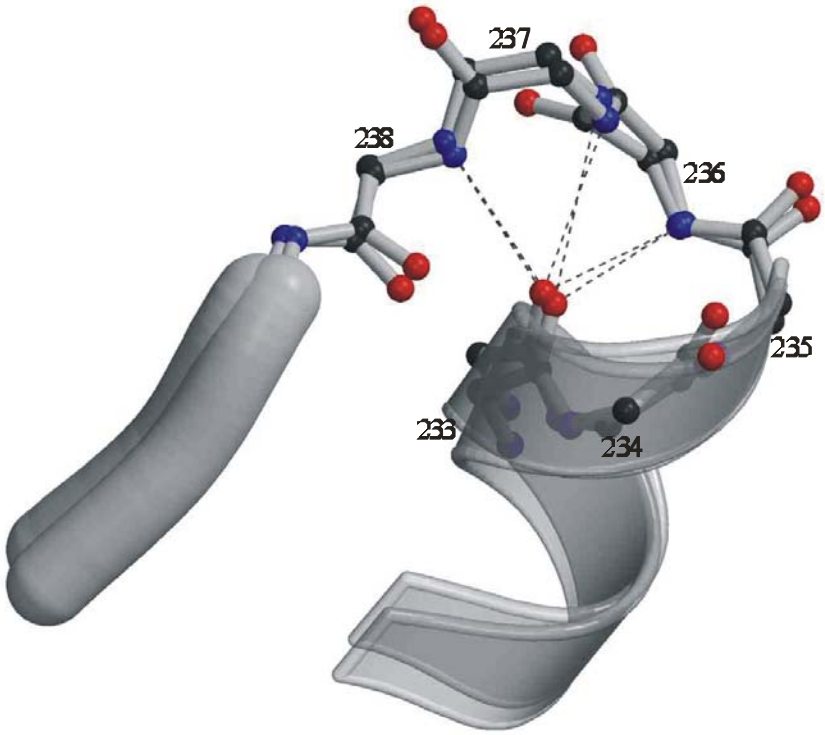


Fig. 3.22 Superimposed α-turn motif from TGEV and HCoV M^{pro}. Side chains are omitted for clarity.

3.3.4.5 Schellman motif

Capping the unsatisfied hydrogen-bonding potential of the carbonyl groups at the C-terminus of an α-helix is an important requirement for its stability. Schellman (1980) reported that about one-third of helices in proteins end in a residue with left-handed (α_L) conformation having a residue with positive angles (ϕ, ψ) at position C' and two main-chain/main-chain hydrogen bonds between residues C3 with C'' and C2 with C'. This has been called the 'Schellman motif'. It was first seen by Watson (1969) at the C-terminus of the H-helix of myoglobin. It is usually favored by glycine residues that have extended freedom of backbone allowing for positive ϕ angles and α_L conformation. It satisfies two successive backbone CO groups while turning the axis of the chain in a direction that terminates the helix – permitting a large exposure of the main chain CO group (Richardson & Richardson, 1988). In TGEV M^{pro}, the B-helix is terminated by such a motif: A210 (C3)-L211 (C2)-I212 (C1)-N213 (C_{cap})-G214 (C')-E215 (C'') (Fig. 3.23). The conformational details of the motif are given below (Table 3.10). According to Aurora *et al*. (1994), the helix capping motifs (Schellman motifs and the α-turns as found involving G255, Section 3.3.4.4) have two 'facets' – one helical and the other not. The helical facet rigidifies the final helical turn, by hydrogen bonding back to the main-chain CO, whereas the other facet terminates the helix and directs the polypeptide chain along a new trajectory. It is likely that such a motif substantially adds to the stability of the overall structure.

The residues in the equivalent position (210-215) in HCoV M^{pro} do not display a Schellman motif. A detailed look at the electron density revealed that its poor quality disallowed any definitive model building. The general direction of the polypeptide chain starting from 210 to 216 is similar in both TGEV M^{pro} and HCoV M^{pro}, except that residues N213 and G214 had particularly ill-defined density in HCoV M^{pro} indicating that presence of this motif cannot be totally ruled out.

Table 3.10 Example for Schellman motif from TGEV M^{pro} chain A

Residue	CO(*i*)...NH (*i*+5) (Å)	CO(*i*+*1*)... NH(*i*+*4*) (Å)	ϕ_{i+1} (°)	ψ_{i+1} (°)	ϕ_{i+2} (°)	ψ_{i+2} (°)	ϕ_{i+3} (°)	ψ_{i+3} (°)	ϕ_{i+4} (°)	ψ_{i+4} (°)
A210-L211-I212-N213-G214-E215	3.03	2.88	-59.04	-58.29	-55.91	-38.22	-90.73	11.19	105.06	-2.60

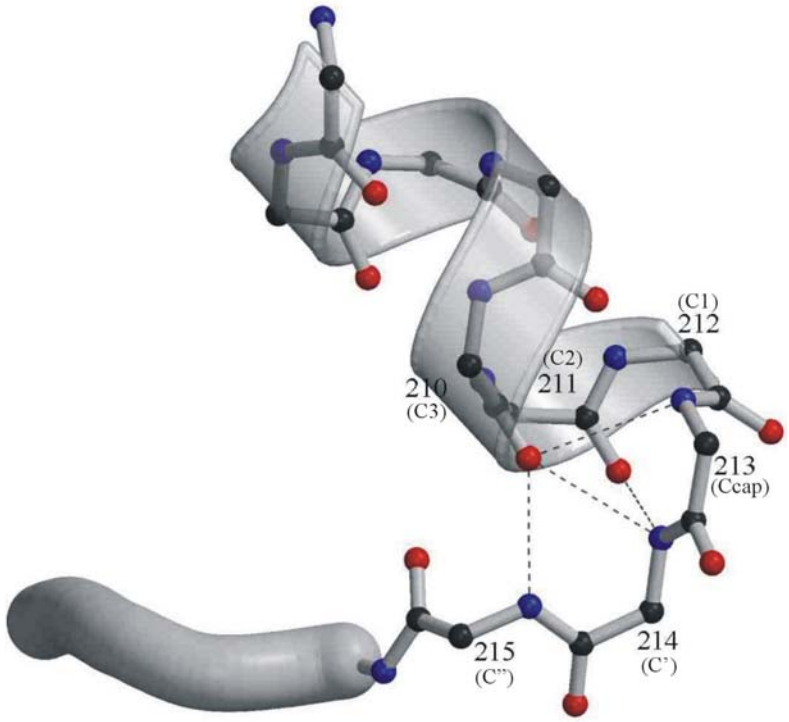


Fig. 3.23 A Schellman motif found in the TGEV M^{pro} crystal structure. For clarity, the side chains are omitted.

3.3.4.6 β-Bulges

Like tight turns, bulges can affect the directionality of the polypeptide chain, but in a much less drastic manner (Richardson *et al.*, 1978). Compared to regular β structures, a β-bulge puts the usual alternation of side-chain direction out of register in one of the strands, thereby introducing a slight bend in the β-sheet. β-bulges can be classified into several different types,

the most common of which is the '*classic*' β-bulge. It is a region between two consecutive β-type hydrogen bonds which includes two residues (position 1 and 2) on one strand opposite a single residue (position x) on the other strand. Residue 1 is in approximately α-helical conformation (averaging ϕ_1 = -100°, ψ_1 = -45°) and residue 2 and x are in approximately normal β conformation (averaging ϕ_2 = -140°, ψ_2 =160°, and ϕ_x = -100°, ψ_x = 130°). There are three classic β-bulges each in the TGEV and HCoV M^{pro} structures conforming to these features (Table 3.11). In TGEV M^{pro}, strand eI involving residues V77 and S78 makes a bulge with K89 (position x) in strand fI. The second bulge is found at G122-S123 in strand cII, and its opposite strand is bII with A115 as the hydrogen-bonding partner. The third β-bulge is at H163-L164 of strand eII, which is antiparallel to the fII strand and involved in hydrogen bonding with G172. In HCoV M^{pro}, the β-bulges are observed at the same position as in TGEV M^{pro} (Fig.3.24; Table 3.11).

Table 3.11 Classic β-bulges found in the A subunit of TGEV and in HCoV M^{pros}

Residues	ϕ_1 (°)	ψ_1 (°)	ϕ_2 (°)	ψ_2 (°)	ϕ_x(°)	ψ_x (°)
TGEV M^{pro}						
V77-S78; K89	-92.79	-25.53	-165.14	162.92	-80.07	131.85
G122-S123; A115	-116.27	-40.02	-156.51	160.01	-77.31	124.41
H163-L164; G172	-117.71	-50.00	-159.90	174.57	-122.44	134.65
HCoV M^{pro}						
V77-G78; K89	-112.92	-18.03	-172.83	161.56	-90.07	117.71
Q122-G123; A115	-117.24	-20-09	-171.86	162.18	-73.18	117.90
Q163-I164; V172	-124.87	-49.23	-144.91	163.96	-134.76	136.36

β-bulges are suited to fill in a protein structure. The β-bulges at 77-78 and 122-123 are located at a junction where the strand leaves the β-sheet in a different direction. His162-His163 are involved in substrate binding and the bulge at this position provides a bend in the β sheet which helps the overall structure to fit together better and orient side chains in the needed direction such that they are appropriately placed at the binding site. The bulges can also provide a mechanism for accommodating a single-residue insertion or deletion mutation without totally disrupting the β sheet.

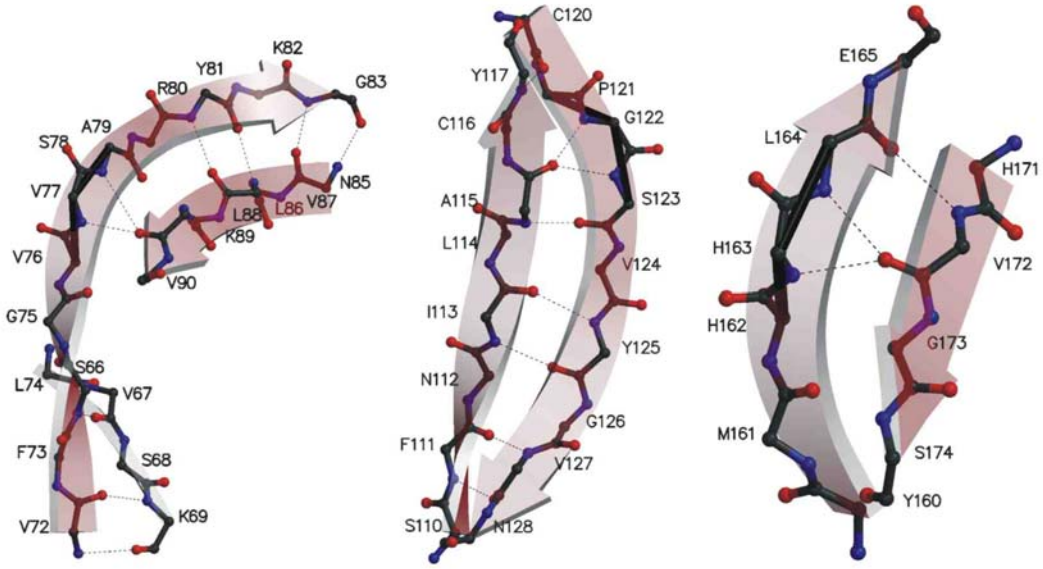


Fig. 3.24 Three β-bulge examples from the TGEV M^{pro} structure. Positions 77 (left), 122 (middle) and 163 (right) also embed the bulge conformation in HCoV M^{pro}. Side chains are omitted for clarity. Thick sticks through the C_α positions of the bulge residues indicate the bulge region.

3.3.4.7 Hydrogen bonding

***TGEV M^{pro}*:** Regular secondary structure accounts for about ~70% of the polypeptide chain. These structural elements are stabilized mainly through hydrogen bonds between backbone N and O atoms. Main-chain-side-chain interactions are fewer in number compared to main-chain-main-chain interactions, as in most other proteins. This observation is consistent with the fact that most polar side chains are distributed on the outer surface of the protein molecule, while a considerable fraction of the main-chain atoms are buried in the interior. Most of the main-chain-side-chain hydrogen bonds are short-range interactions occurring within four residues along the polypeptide chain. The 717 hydrogen bonds were counted for the calculation of mean value distances in α-helices is 3.10 (± 0.14) Å in all six molecules. The short 3_{10} helices gave the value 3.11 (± 0.139) Å for 85 H-bonds. The H-bonds counted for 786 β-sheets, gave mean value of 3.04 (± 0.09) Å for all molecules. The 244 hydrogen bonds in β-turns, on an average, are 3.07 (±0.11) Å long.

***HCoV M^{pro}*:** The secondary structure distribution is similar to that of TGEV M^{pro}. The mean value of 223 calculated hydrogen-bond distances in α-helices is 3.08 (± 0.14) Å for the dimer of HCoV M^{pro}. For the short 3_{10} helices the value is 3.15 (± 0.17) Å out of 12 H-bonds. The 260 β-sheet-hydrogen bonds gave the mean value of 3.04 (± 0.09) Å for both molecules, and 66 hydrogen bonds in β-turns, on an average, are 3.14 (± 0.17) Å long.

3.3.5 Dynamic aspects

3.3.5.1 Thermal parameters

Electron density maps provide direct evidence for flexibility in protein molecules. The B-factor indicates the true static or dynamic mobility of an atom, it can also indicate where there are errors in model building. The B-factor is given by $B_i = 8\pi^2 U_i^2$, where U_i^2 is the mean square displacement of atom i.

This produces a weighting factor on the contribution of atom i to the Fourier transform by: $\exp(-B_i \sin^2\theta/\lambda^2)$. As U increases, B increases and the contribution of the atom to the scattering is decreased. If atoms are incorrectly built, their B-factors will tend to be higher than those for correctly built atoms nearby.

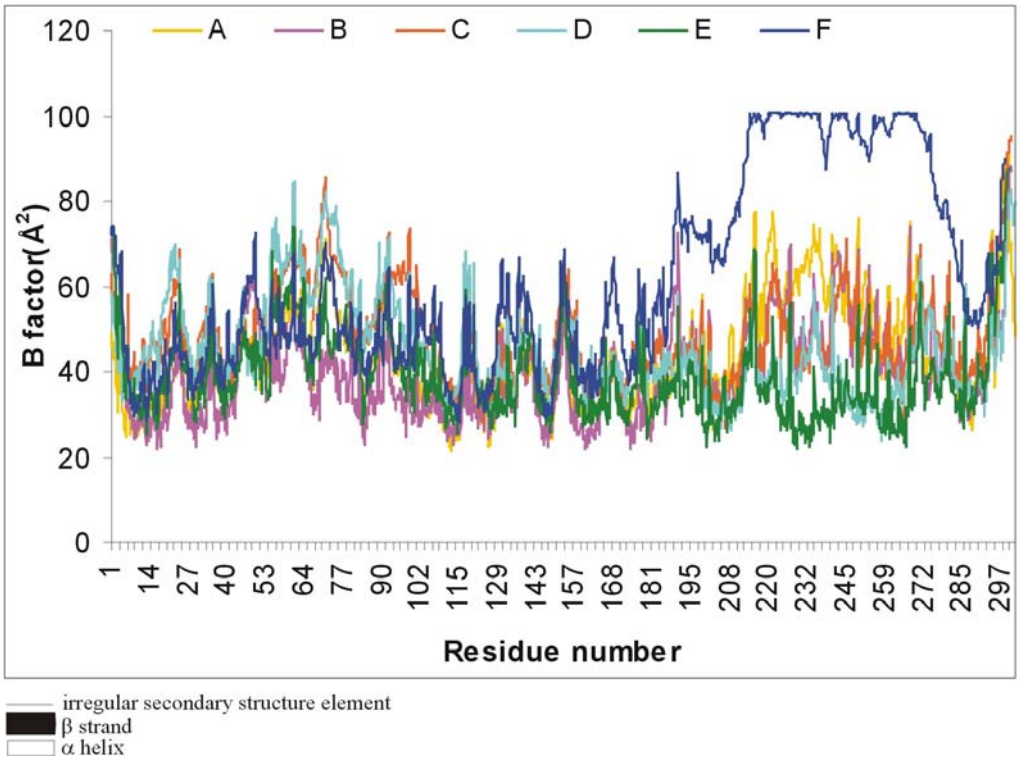
The value of the temperature factor is affected not only by the flexibility of the protein molecule but also by errors in the crystallographic phases and by disorder in the crystal lattice. Thus an interpretation of B values is not straightforward. Lattice disorder and internal flexibility can be distinguished by comparing different parts of the electron density map to determine which variation can be explained by a slightly different packing of rigid molecules in the unit cell, by comparing maps of the same molecules in different crystal lattices or in different environments in the same lattice (*e.g.* when there is noncrystallographic symmetry), or by varying the temperatures of data measurement.

TGEV M^{pro}: Figure 3.25A shows the plot of average B values for main-chain atoms (N,C^{α},C,O) as a function of the residue number. The average temperature factor is 46.1 $Å^2$ for the main-chain atoms, 47.2 $Å^2$ for side-chain atoms and 47.0 $Å^2$ for all protein atoms. The domain III of monomer F shows large differences in B values (~70 $Å^2$), compared to the other monomers A-E.

As evident from the three-dimensional distribution of B-factors (Fig. 3.25B), monomers A, B, and E have relatively large areas with lower thermal parameters (indicated by blue). The inter-monomer loop regions also have lower temperature factor values compared to other exposed regions. Monomer F, especially its third domain, has high temperature factors, albeit with limited influence on other protein regions in the contiguous space. In both TGEV and HCoV proteinases, it is found that the buried residues (ASA < 0.07 $Å^2$) have lower temperature (~39 and ~19 $Å^2$, respectively) factors than the exposed ones (~43 and 28 $Å^2$, respectively).

HCoV M^{pro}: Figure 3.26A shows the B-factor plot of the HCoV M^{pro} structure. Except for the residues near the two termini and in irregular secondary structure elements (loops, bends, 3_{10}-helices, turns, β-ladder, etc), all residues have comparable B values in the two molecules and show the same pattern of variation along the chain. The B values peak at the turns and loops exposed to solvent, while making troughs at the ends of the strands that form the core of the protein molecule (Fig. 3.26B). The average temperature factors are 27.03 $Å^2$ for the main-chain atoms, 27.8 $Å^2$ for side-chain atoms and 27.08 $Å^2$ for all protein atoms. The loop-helix segments 213-231 and 240-260 are the regions with significantly increased B-factor values, although these regions have well defined electron density.

A



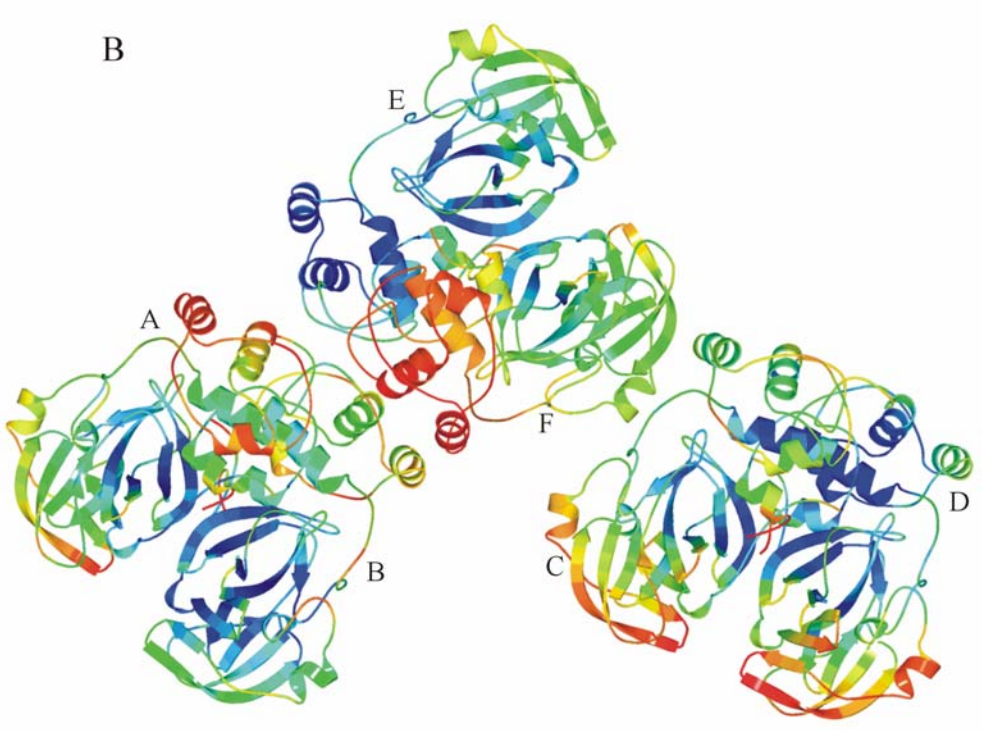
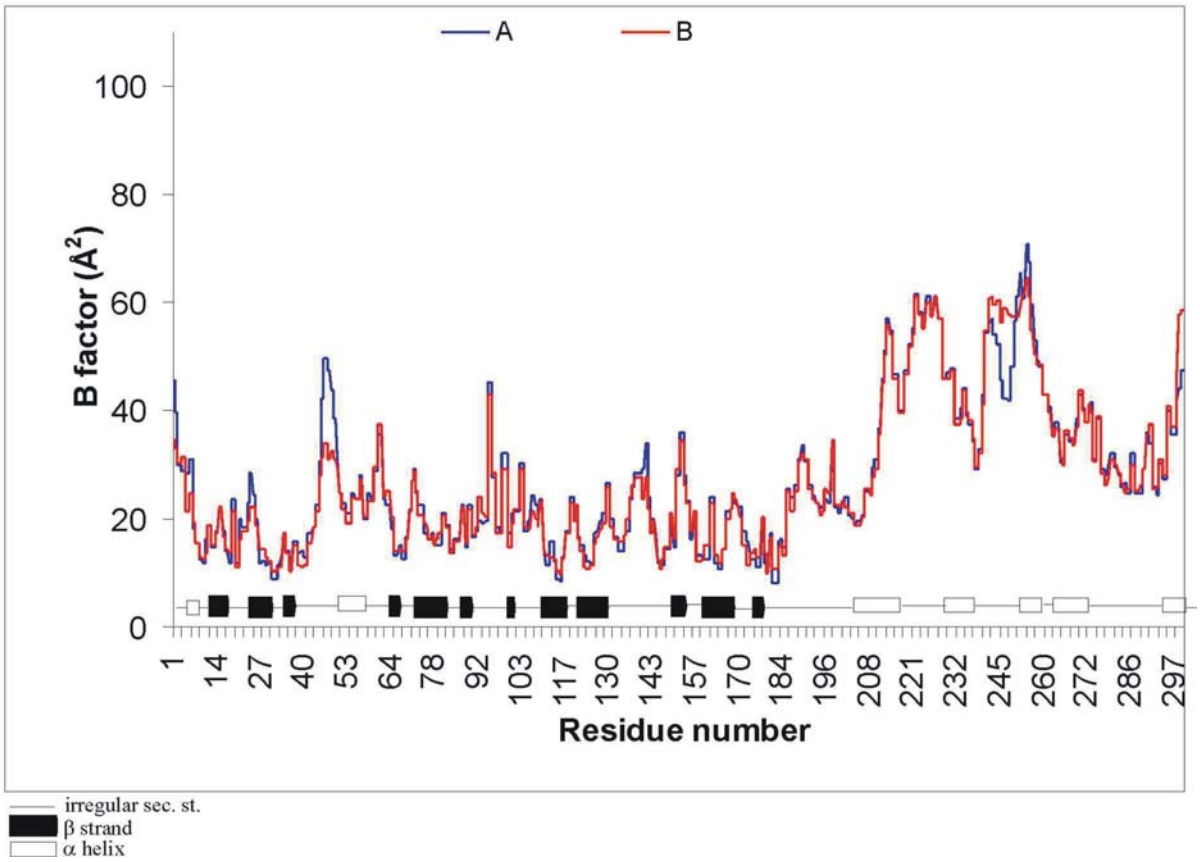


Fig. 3.25 **A.** Average temperature factor for each of the six monomers of TGEV M^{pro}. The average B-factor of the C-terminal domain of monomer F is distinctively higher than that of the other monomers. The B-factor limit (program CNS) here is 100 $Å^2$. **B.** Variation of the temperature factors in the three-dimensional structure. The color code ranges from blue to red indicating lower to higher temperature factors. Monomers A–F are labeled.

A



B

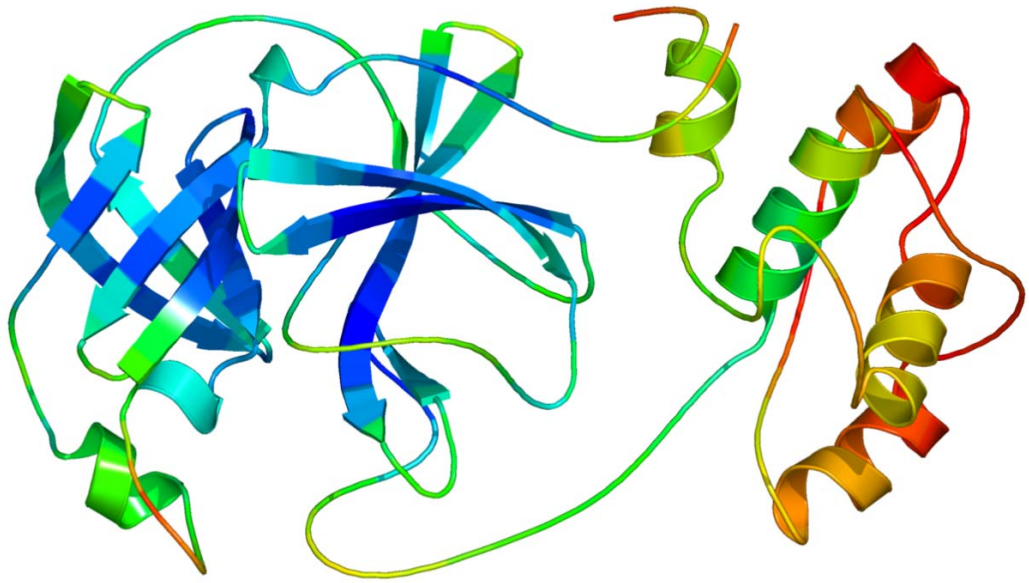


Fig. 3.26 **A.** B-factor plots for the HCoV M^{pro} crystal structure showing the flexible C-terminal domain with relatively higher thermal factors than for the other two domains.

B. Variation of the temperature factor in monomer A. Color codes are as in Figure 3.23. Monomer B has a similar B-factor distribution (see panel A).

3.3.5.2 Solvent Structure

Water molecules contribute to the overall stability of proteins and to the characteristics of the surface (Caspar & Badger, 1991; Herron *et al.*, 1994; Karplus & Faerman 1994; Zhang & Matthews, 1994). They influence protein folding, maintain the native structure and can be critical for the function, e.g. enzyme activity (Rupley *et. al*., 1980). In the TGEV M^{pro} asymmetric unit, there is clear electron density for at least 1006 water molecules while 221 sites were located for the HCoV M^{pro} structure for water molecules. Most of these water molecules interact with polar portions of the amino-acid residues on the surface of the protein, whereas only a few are buried within the protein core (see appendix, Table 6.3A) (Lee & Richards, 1971; Rashin *et al*., 1986; Tilton *et al*., 1986). The water molecules form an extensive network of hydrogen-bonding interactions in both crystals. In TGEV M^{pro}, the buried water molecules, W11, W14, W15, W28, W204, W258 and W321 are related by NCS. They superpose on each other and are located in the active-site cavity (Fig. 3.31) of each of the six monomers. Many of the water molecules are well ordered, bind firmly through three or four hydrogen bonds and are located in surface pockets and crevices. The water molecules related by non-crystallographic symmetry have comparable temperature factors and occupancies.

All the water molecules except W10, W291, W499, W800, W884, W909, W955, W983, W984, and W1002 form at least one hydrogen bond to a protein atom directly, and thus occur in the first hydration shell. The water molecules that are further than 3.3 Å from the protein have higher B values and form hydrogen bonds to the protein surface through other water molecules.

3.3.5.3 Electrostatic surface

Surface representations are an important tool for representing molecules and for computing inter-molecular interactions. Various molecular properties, such as atomic charge, electrostatic potential, hydrophobicity, polarizability, etc. strongly influence the way in which molecules interact with one another. In order to represent this, the information is displayed onto the molecular surface color codes, e.g., in the surface color-coded by electrostatic potential, the red areas (where the surface is negative) are interpreted as areas, which will most favorably interact with another molecular surface, which is blue (where the surface is positive). These kinds of representations are used to understand and predict how molecules will interact with one another, dock with one another and even to design molecules that will optimally bind to another molecule.

The overall surface charge for both TGEV M^{pro} and HCoV M^{pro} is largely neutral (Fig.3.27A and 3.27B); however, HCoV M^{pro} has even fewer charged patches than TGEV M^{pro}. Even the active site groove bears only some negative potential.

The M^{pro} molecules were rotated to compare and align with the KFDRI region of the HAV 3C proteinase. The interdomain connection between I and II, is a surface that has been demonstrated to be involved in RNA recognition in picornavirus 3C proteinases (Bergmann *et al.*, 1997). In contrast to the latter proteinases, very few charged residues contribute to the corresponding surface of M^{pro}. Further investigations for RNA-binding site continued with domain III. In domain III, there is no RNA binding site detectable, as the surface does not contain any particular patches of the basic residues. Domain III was individually checked for the overall charge pattern, which is again the same as that of the domain I and II. No distinct patches of basic amino acid residues were found.

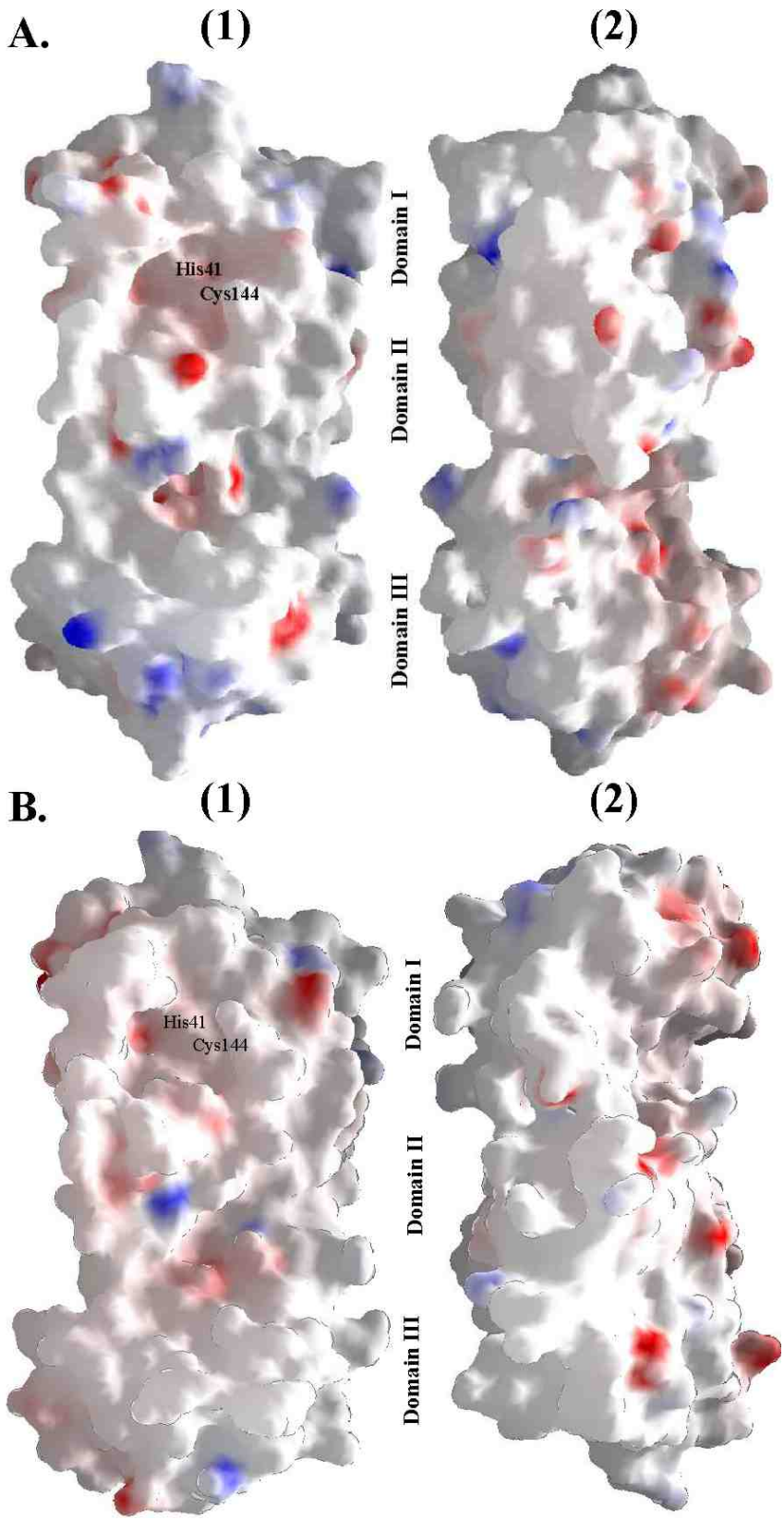


Fig. 3.27 Surface representation (GRASP, Nicholls *et al.*, 1991) of **A**., TGEV M^{pro} and **B.,** HCoV M^{pro}. **1**. Surface containing the active-site cleft; the catalytic residues Cys144 and His41 are indicated. **2**. The M^{pro} molecule is rotated by approx. -90° relative to the view shown in **1** to compare and align with the KFDRI region of the HAV 3C proteinase. Positive and negative surface charges are colored blue and red, respectively.

3.3.6 Crystal Packing

The area of the protein surface involved in packing contacts is generally small and its amino acid composition is indistinguishable from that of the protein surface accessible to the solvent. Crystal packing provides examples of nonspecific protein-protein interactions, which can be compared to biologically relevant ones. The number of protein chains and protein segments involved in a crystal-packing contact is variable. The fraction of the protein surface involved in crystal contacts is very variable and independent of the number of packing contacts. A study by Janin & Rodier (1995) mainly focused on the geometrical features of the thermal motion at the crystal-packing interface and on the relation between surface burial and symmetry of the crystal. The number of crystal contacts and the percentage of the protein surface buried by them, decreases as the percentage of the solvent content increases (Matthews, 1968). Similarly, the temperature factors at the crystal packing contacts decreases as the area of the surface involved in the contact increases (Carugo & Argos, 1997). These general points have been confirmed during the analysis of the TGEV/HCoV M^{pro} structures.

The lattice arrangement of the six TGEV M^{pro} monomers is shown in Figure 3.28 A, B and HCoV M^{pro} dimer in Figure 3.28 C, D. Analysis of contacts between symmetry-related TGEV M^{pro} molecules shows that the molecules F-C and E-B have more number of contacts as compared to the other monomers. The A and D monomers are less involved in the crystal contacts. Most of the regions involved in subunit association have irregular secondary structure, namely loop regions, bends, turns and 3_{10} helices. Helix D and E of TGEV M^{pro} monomers pack against each other in the dimer, and so do β-strands aI and fI. The lattice is stabilized by quite a few strong electrostatic interactions (see appendix, Table 6.2.1-6.2.4). An elaborate web of hydrogen bonds supplements these interactions. There are at least 80 strong hydrogen bonds involved in crystal packing among the symmetry-related molecules. The interfaces between the molecules are essentially hydrophilic in nature with no residues being disordered.

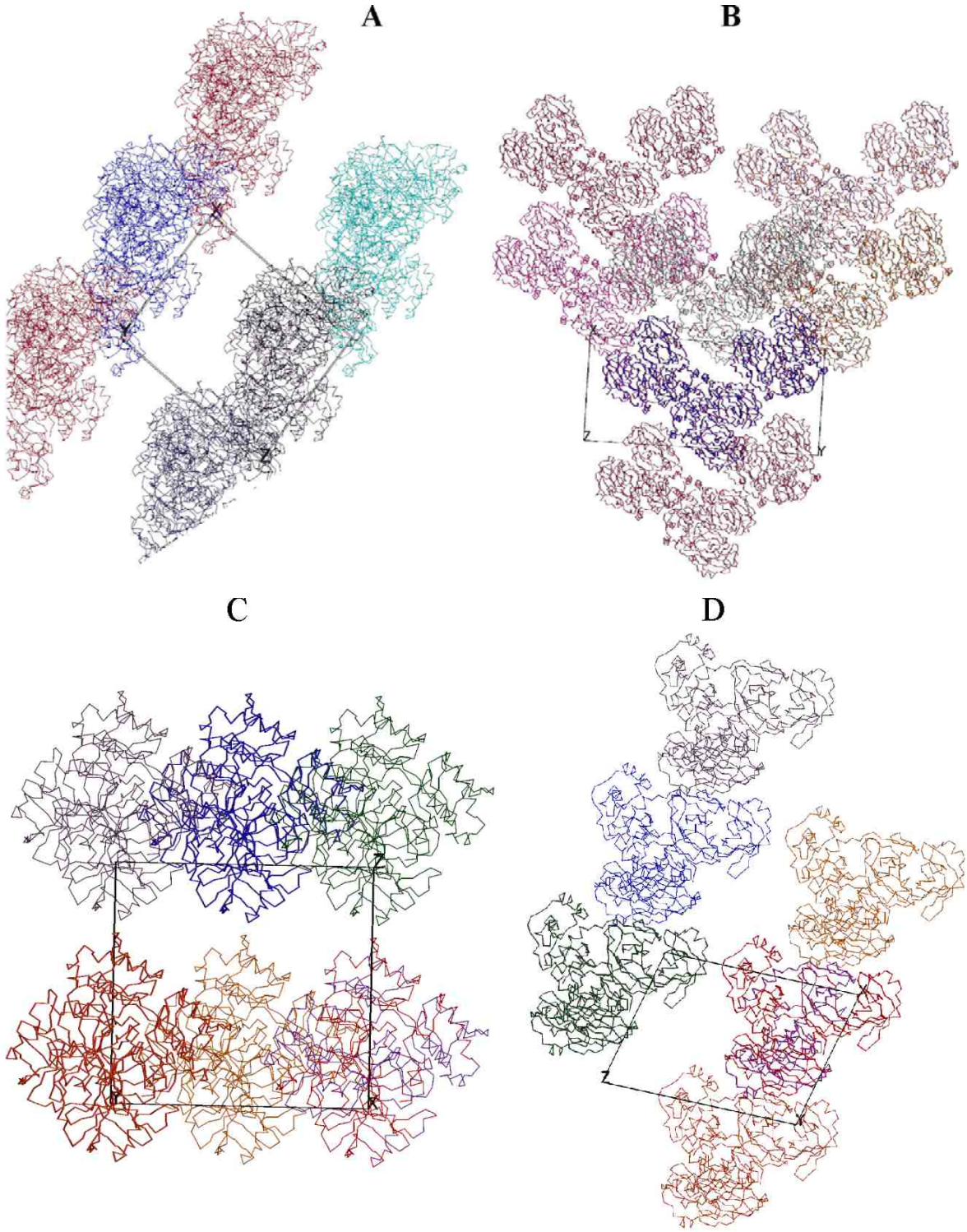


Fig 3.28 Crystal packing. TGEV M^{pro}: **A.** View along the crystallographic y-axis **B.** view along the z-axis, whereas HCoV M^{pro}: **C.** view along the y-axis and **D.** along the z-axis. All copies of the TGEV M^{pro} and HCoV M^{pro} are shown in individual colors.

3.3.7 Bound crystallization additives

Additional electron density was found in both the TGEV and HCoV M^{pro} structures, on the surface and in the active sites. In TGEV M^{pro}, nearly spherical densities close to arginines 19, 61, 130, 216, 275, and 294 were interpreted as sulfate ions. They make reasonable hydrogen bonds with the guanidinium groups of these arginines and with adjacent main-chain nitrogens. Some of the sulfates are half, some are fully occupied.

In our electron density maps, part of the S2 subsite (of all six copies of the monomer) harbours extra electron density that is interpreted as an MPD molecule from the crystallization medium (Fig. 3.30), (the crystallization precipitant contained additives such as MPD and dioxane). 6 MPD and nine dioxane molecules were modeled into difference density in the TGEV M^{pro} crystal structures. In the case of both HCoV M^{pro} monomers, a dioxane molecule was built into the S2 subsite (Fig. 3.29) (see appendix 6.5).

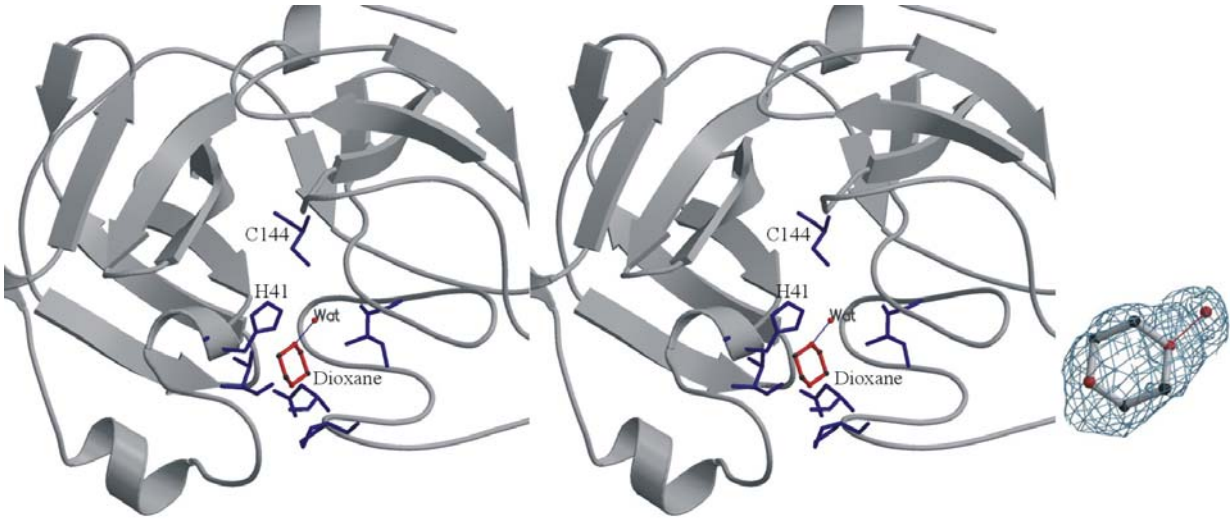


Fig. 3.29 Stereo figure showing the environment of a dioxane molecule bound in the substrate-binding site of HCoV M^{pro}. Active site residues are labeled; the distance between the oxygen atom of dioxane and the water molecule is 2.17 Å. Extreme right: electron density map (2|Fo|-|Fc|) for the dioxane molecule (2.7σ above the mean).

The MPD molecule binds near the presumed S2 and S3 subsites of the substrate-binding site, between the two β-barrel core domains. This MPD molecule is also interacting with the long loop (residues 184 to 199) connecting domains II and III; this loop is involved in substrate binding (Fig. 3.20 & 3.30).

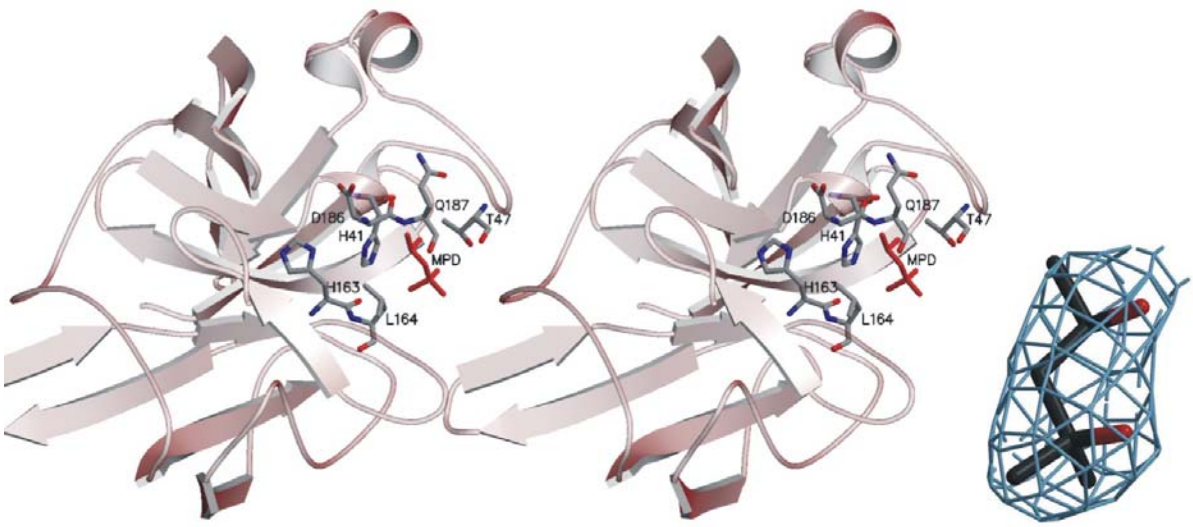


Fig. 3.30 The first two domains of TGEV M^{pro} where an *R*-MPD molecule binds to the substrate-binding site and interacts with two additional residues from the loop connecting domains II and III. (The loop is shown in figure 3.13, 3.20). Extreme right: an MPD molecule bound in electron density (2|Fo|-|Fc|), at 2.7σ above the mean.

The C1 atom of the MPD contacts the C^{β} of Leu164; O2 forms a hydrogen bond with the main-chain oxygen of Thr47 and C5 stacks with the peptide bond connecting Asp186 and Gln187. The protons attached to C4 can possibly interact with the imidazole π electrons of His41.

The physicochemical properties of the substrate recognition sites of many proteinases suggest that, although they are not always hydrophobic, they may have hydrophobic subsites, which can be potential binding sites for MPD and other solvent molecule(s). The role of MPD in protein crystallization is stabilization of proteins by filling up voids and cavities on the protein surface rather than glue for binding proteins together in a crystal lattice (Anand *et al*., 2002b). This is seen in the TGEV M^{pro} crystal structure where MPD does not take part in crystal contacts, but lies freely in the active-site groove.

3.4 Functional implications of the structures

3.4.1 Catalytic system of M^{pro}

The active site of the coronavirus M^{pro} (Fig. 3.31) is similar to those of the picornavirus 3C proteinases, as had been predicted earlier (Gorbalenya *et al.*, 1989b). The mutual arrangement of the nucleophilic Cys144 and the general acid-base catalyst His41 of TGEV M^{pro} is identical to that of the HAV $3C^{pro}$ residues Cys172 and His44 and to the residues Ser195 and His57 of

chymotrypsin. In TGEV M^{pro}, the average distance between the sulfur atom of Cys144 and the $N^{\varepsilon 2}$ of His41 is 4.05 (± 0.04) Å, *i.e*. longer than the corresponding Cys-to-His distances in HAV $3C^{pro}$ (3.92 Å; Bergmann *et al.*, 1997), poliovirus (PV) $3C^{pro}$ (3.4 Å; Mosimann *et al.*, 1997), and papain (3.65 Å; Kamphuis *et al.*, 1984) (Fig. 3.31A and 3.32B). In contrast to papain, but in agreement with the picornavirus 3C proteinases, the sulfur atom is in the plane of the histidine imidazole. There are clear indications from the difference Fourier synthesis that Cys144 is oxidized (Fig. 3.32A), at least to the stage of the sulfinic acid, $-SO_2^-$, and probably to the sulfonic acid, $-SO_3^-$, in all six copies of TGEV M^{pro} in the crystal. Such oxidation could occur during the time required for crystallization or during X-ray data collection and would lead to inactivation of the enzyme.

The active-site residues are conserved between TGEV and HCoV M^{pro} (Fig. 3.31A & B); in HCoV M^{pro}, the distance between the active-site Cys144 and His41 (3.75 ± 0.04) Å is shorter than found in the TGEV M^{pro}. The active-site cysteine residues in both M^{pro}s are oxidized. In HCoV M^{pro}, the difference electron density at the active-site cysteine of both the molecules in the asymmetric unit could be interpreted as a sulfenic acid ($-SO^-$). Refinement of the oxidized Cys derivatives was not successful.

It is generally assumed that the native state of the active site of papain-like cysteine proteinases is a thiolate-imidazolium ion pair formed by cysteine and histidine residues (Polgár, 1974). In proteinases of the papain family, an asparagine is the third member of the catalytic triad. Chymotrypsin and other members of this serine proteinase family have a catalytic triad consisting of Ser195...His57...Asp102. In HAV $3C^{pro}$, Asp84 is present at the required position, although its side chain points away from His44, making its role disputable (Malcolm, 1995; Bergmann *et al.*, 1997). PV $3C^{pro}$, human rhinovirus (HRV) $3C^{pro}$, and HRV $2A^{pro}$ have a glutamate or aspartate in the proper orientation to accept a hydrogen bond from the active-site histidine (Matthews *et al.*, 1994; Mosimann *et al.*, 1997; Petersen *et al.*, 1999). In contrast, both TGEV and HCoV M^{pro}s have Val84 in the corresponding position, with its side chain pointing away from the catalytic site (Fig. 3.31 A, B and 3.32B). In both structures, a buried water molecule is found in the place that would normally be occupied by the side chain of the third member of the catalytic triad. This water molecule makes hydrogen bonds to His41 $N^{\delta 1}$, His163 $N^{\delta 1}$, and Asp186 $O^{\delta 1}$ (Fig. 3.31).

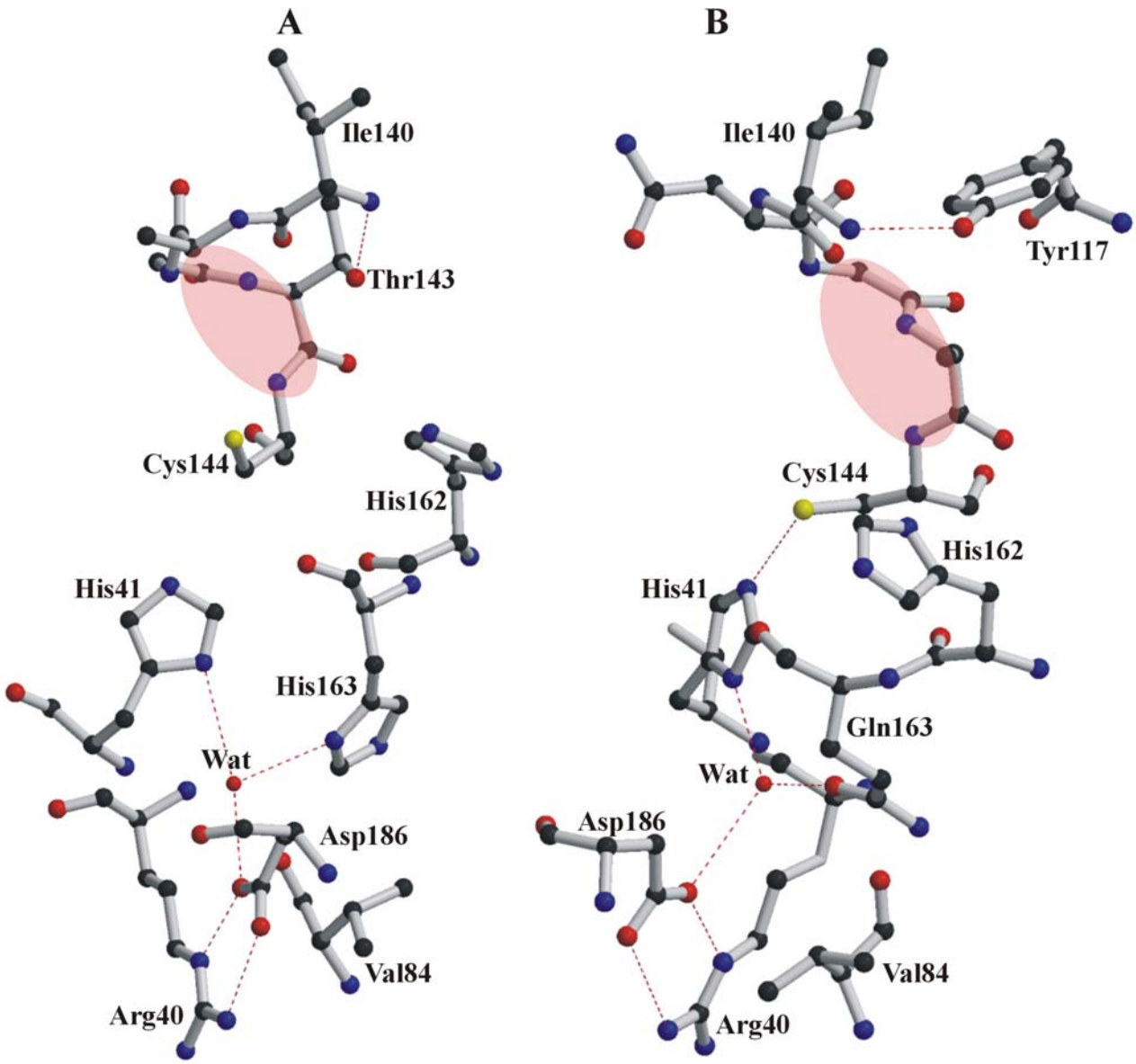


Fig. 3.31 The active site of coronavirus M^{pro}. **A.** The catalytic Cys144 and His41 residues of TGEV M^{pro} are shown. The region forming the oxyanion hole (main-chain amides of Gly142, Thr143, Cys144) is highlighted in pink. The water molecule, which occupies a position equivalent to that of the catalytic Asp of serine proteinases, is shown together with its hydrogen-bonding partners His41, His163 and Asp186. **B.** Similar diagram for HCoV M^{pro}. There is Ala143 instead of Thr143 in the oxyanion hole. The main chain amide of Ile140 is strongly interacting with OH of Tyr117, thus stabilizing the oxyanion hole. There is Gln at position 163 instead of His. The active-site geometry is similar to that of TGEV M^{pro}. Asp186 cannot take the position of the water molecule through rotation about its Cα-Cβ bond. It is making a strong salt-bridge with Arg40.

His163 is not conserved among coronavirus main proteinases and its substitution by Leu (M^{pro}-H163L) had no significant effect on the proteolytic activity in the standard peptide assay (see Section 3.4.8), as compared to the activity of the wild-type M^{pro} (Ziebuhr *et al*., 2000) (Table 3.12). Asp186 makes a salt bridge to Arg40 that appears to be required to maintain the active site geometry – both residues are absolutely conserved among coronaviruses. Through this (and other) interaction(s), the polypeptide segment 184 – 199, which connects domains II and III and is probably involved in substrate binding, is held in the proper position. Taken together, the data contradict a *direct* involvement of His163 or Asp186 in catalysis, making both the TGEV M^{pro} and HCoV M^{pro} clear case, of viral cysteine proteinases employing only a catalytic dyad.

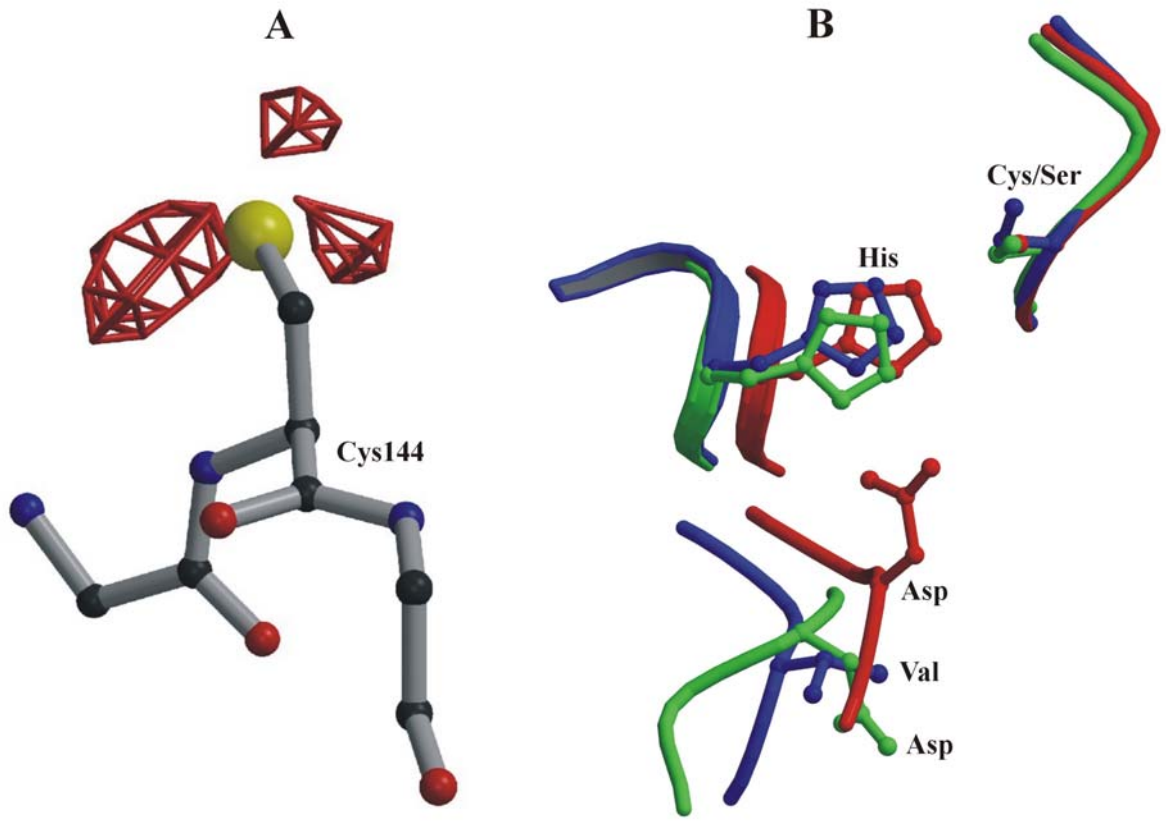


Fig. 3.32 **A.** The catalytic Cys144 in TGEV M^{pro} is shown along with the difference electron density (|Fo|-|Fc| at 3.0 σ above the mean; red) for the oxidized state – indicating three oxygen atoms bound to the sulfur.

B. Superposition of the active-site residues, including the third catalytic residue (Asp102) of chymotrypsin (shown in red), amino-acid residue Asp84 of HAV $3C^{pro}$ (side chain oriented differently; green) and Val84 (blue) of TGEV M^{pro} are shown.

3.4.2 Transition state stabilization

It is generally accepted that substrate hydrolysis by cysteine proteinases occurs through a covalent tetrahedral intermediate resulting from attack of the active-site nucleophile on the carbonyl carbon of the scissile bond. In all known cysteine and serine proteinases, distortion of the carbonyl of the substrate is stabilized by strong hydrogen bonds between the developing oxyanion and amide groups of the enzyme. This so-called 'oxyanion hole' is also found in both TGEV M^{pro} and HCoV M^{pro}. It is made up by the main-chain amides of Gly142, Thr143 (Ala143 in HCoV M^{pro}), and Cys144 (Fig. 3.31 A, B). The glycine is in the *i+2* position of a type II reverse turn, which immediately precedes a second type II turn carrying the nucleophile. In TGEV M^{pro}, Thr143 plays a pivotal role in stabilizing the two connected turns by properly orienting the main-chain amides that donate the hydrogen bonds to the oxyanion. The side-chain hydroxyl group of this residue accepts a 3.09 (±0.09) Å hydrogen bond from the main-chain amide of Ile140, and it donates a 2.94 (±0.19) Å H-bond to the main-chain carbonyl of the same residue. The threonine side-chain is absent in HCoV M^{pro} (Ala143). Here, the proximal Tyr117 OH stabilizes the two connected turns. This is similar in the picornavirus proteinases or chymotrypsin, where the architecture of the oxyanion hole is very similar to the coronaviruses but employs different stabilizing interactions. The distinct hydrogen-bonding pattern is similar to that in bacterial chymotrypsin-like proteinases such as α-lytic proteinase (αLP; Fujinaga *et al*., 1985) and *Streptomyces griseus* proteinase A (SGPA) (James *et. al*., 1980).

Chymotrypsin-like serine proteinases have the sequence Gly(193)-Asp(194)-Ser(195)-Gly(196) around the nucleophile. Asp194 forms a hydrogen bond both to the main-chain N at position -3 from the nucleophile in its own reverse turn and to β strands topologically identical to aII and bII1 found in HRV2-$2A^{pro}$ (Fig. 3.33). In contrast, a similar hydrogen-bonding pattern is not found in the $3C^{pro}$s of PV and HAV. In the PV $3C^{pro}$, the residue preceding the nucleophile is Gln, whereas in HAV $3C^{pro}$ it is Met. In both cases, there are no interactions with strands aII and bII1. Replacement of the active site Cys172 in HAV $3C^{pro}$ by Ala causes the oxyanion hole to collapse (Allaire *et al.*, 1994). This suggests that the thiolate ion of the proposed imidazolium-thiolate ion pair stabilizes the oxyanion hole by electrostatic interaction, forcing the main-chain carbonyl oxygens from the two residues preceding the nucleophile away from the thiolate as seen in HRV2-$2A^{pro}$ (Petersen *et. al*., 1999). The structure of the oxyanion hole in monomer A of TGEV proteinase is shown in Figure 3.31A

and in HCoV M^{pro} in Figure 3.31B. The active-site Cys144 is oxidized but nevertheless takes part in making the oxyanion hole.

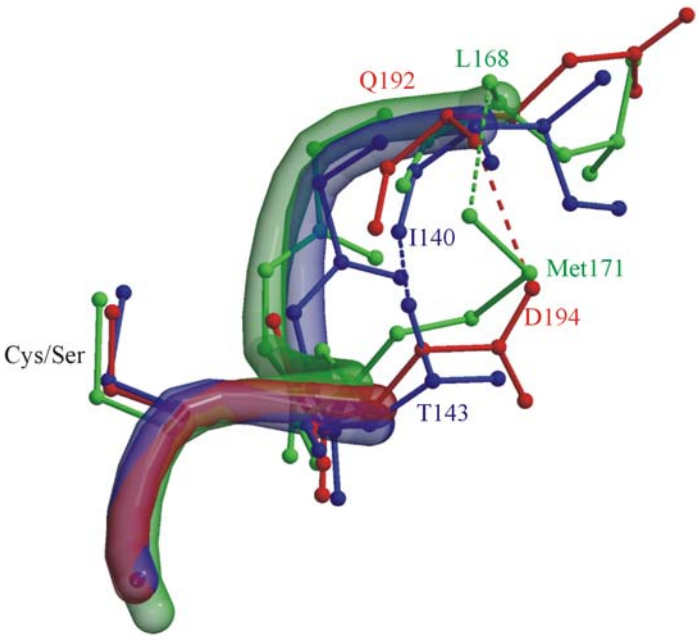


Fig. 3.33 Superposition of oxyanion holes of TGEV M^{pro} (blue), HAV $3C^{pro}$ (green) and chymotrypsin (red) is shown. Labels are colored accordingly.

One of the oxygens bound to the sulfur atom makes a H-bond with the main-chain amide of Cys144. The cysteine could have been oxidized in solution as well as in the crystal. DTT was used during protein purification. Furthermore, 5mM DTT was used in crystallization, but still the Cys144 is oxidized. This could be due to a combination of various factors: (i) DTT degrades with time as the crystallization proceeds for longer period (~10 days), (ii) DTT may not have been used in sufficient amounts, (iii) Cys144 is exposed to the solvent, (iv) radiation damage during data collection may also cause oxidation.

3.4.3 TGEV M^{pro} - TLCK complex

At present, the available experimental evidence for the catalytic mechanism of coronavirus M^{pro}s points to a catalytic dyad (Anand *et.al*., 2002a). For a better insight into this, HCoV M^{pro} was tested for sensitivity towards common proteinase inhibitors by Dr. J. Ziebuhr. Six compounds effectively inhibited the M^{pro}-mediated peptide cleavage. These were 3,4-dichloroisocoumarin, phenylmethyl-sulfonyl fluoride (PMSF), PefablocSC [4-(2-Aminoethyl) benzenesulfonly fluoride HCl ($C_8H_{10}NSO_2FHCl$)], TLCK, antipain [(S)-I-carboxy-2-Phenylethy-carbamoyl-L-Arg-L-Val-Arginal, ($C_{27}H_{44}N_{10}O_6$)] and $ZnCl_2$. Inhibition was defined as a reduction in cleavage product formation to less than 15% of the inhibitor-free control reaction. In contrast, TPCK, leupeptin, elastatinal, chymostatin, E-64, pepstatin, and

EDTA had no significant inhibitory effect. Thus despite having a nucleophilic Cys residue which cannot be functionally replaced by Ser (Ziebuhr *et. al*., 1995), the HCoV M^{pro} enzyme can be effectively inhibited by several serine proteinase inhibitors. Importantly however, the inhibitory effect of PMSF can be reversed by dithiothreitol, supporting the presence of a cysteine in the active site.

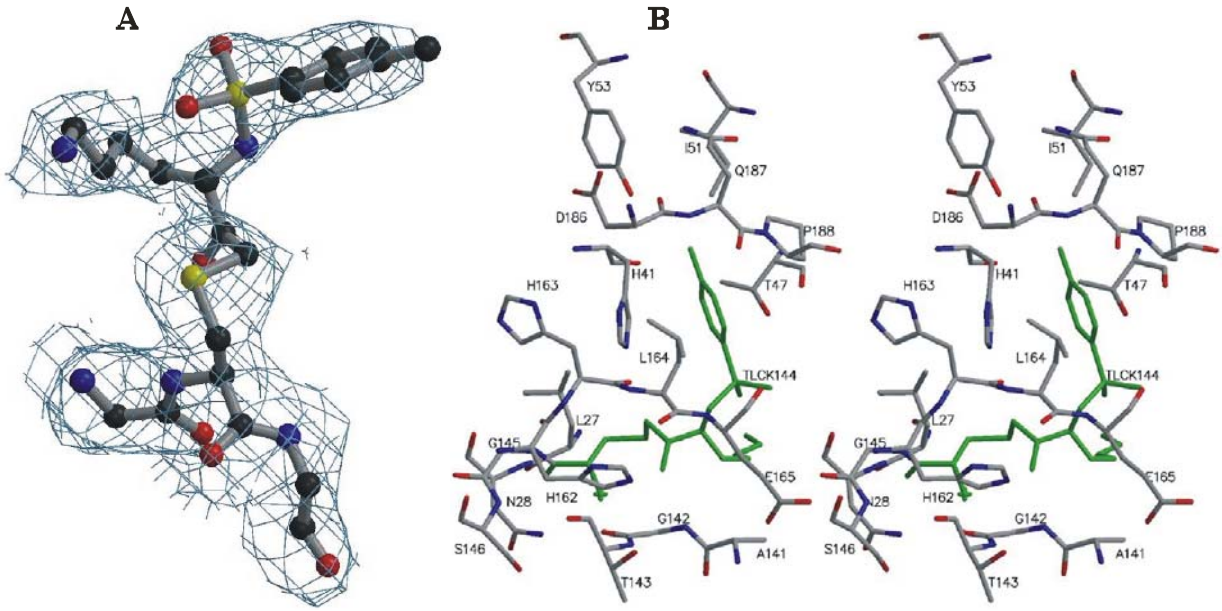


Fig. 3.34 TLCK bound to the active site Cys144. **A.** The 2|Fo|-|Fc| electron density map (2.6 Å resolution of TGEV M^{pro}-TLCK complex), contoured at 1σ above the mean. **B.** Stereo view showing residues interacting with TLCK bound to Cys144.

To study the active-site geometry in more detail and to provide further support for the catalytic role of Cys144, the structure of TGEV M^{pro} in complex with TLCK (Fig.3.34) was determined by molecular replacement and refined to 2.6 Å resolution. The wild-type TGEV M^{pro} crystals were soaked into five fold molar excess of TLCK inhibitor for 14 hr. The TLCK did not change significantly the unit cell dimensions of native TGEV M^{pro}. Both data sets were quite isomorphous that is why the TGEV M^{pro}-TLCK complex structure could be solved by molecular replacement method. The unit cell dimensions of the crystals are a = 72.8 Å, b = 160.1 Å, c = 88.9 Å, β = 94.3° and display space group $P2_1$. In this structure, the distance between the sulfur atom of Cys144 and $N^{\varepsilon 2}$ of His41 is 4.02 Å. The TLCK molecule binds covalently to the active site Cys144 in all monomers of TGEV M^{pro}, as well as to Cys120 and Cys284. These two cysteines are exposed to the solvent channel. Therefore, binding of TLCK failed to provide any significant information. This result led to the design of a substrate-

analog inhibitor, of which the structure was solved in complex with TGEV M^{pro} (see Section 3.4.5).

3.4.4 Substrate binding site

The substrate specificity of TGEV and HCoV M^{pros} resembles that of many other 3C/3C-like proteinases (Kräusslich & Wimmer, 1988; Dougherty & Semler, 1993). The arrangement of the residues in the substrate-binding site of both M^{pro}s is similar, with an rmsd value of 0.65 Å for all atoms (132 target pairs out of 144 C_α pairs) of M^{pro}s residues (Figure 3.18). There are minor variations in the conformation of the side chain for some residues. The P1 position of the substrate is occupied by Gln, and small residues (Ser, Ala, Gly, Cys, or rarely Asn) are found at the P1' position. Coronavirus main proteinases exhibit conservation mainly of Leu in the P2 position of their substrates. In contrast, HAV $3C^{pro}$ prefers Ser/Thr at P2, whereas PV $3C^{pro}$ tolerates a wide variety of amino acids at this position. At P4, Val/Thr/Ser are found in coronavirus M^{pro} substrates, whereas the HAV and PV 3C proteinases favour Leu/Ile/Val and Ala/Val/Pro/Thr, respectively. TGEV deviates slightly from other coronaviruses (including HCoV M^{pro}) in that, in one cleavage site, Met is found at P2, while, in two others, Ile or Lys occurs at P4. One can speculate that these non-canonical cleavage sites are less efficiently processed than the more usual substrates.

In order to visualize potential interactions with the substrate, a pentapeptide representing the P5 – P1 residues of the N-terminal cleavage site of the TGEV M^{pro} (Asn-Ser-Thr-Leu-Gln) was initially modelled into the substrate-binding cleft of the TGEV main proteinase (Fig. 3.35) and later the X-ray structure of TGEV M^{pro} in complex with a P6-P1 peptidyl chloromethyl-ketone (Z-Val-Asn-Ser-Thr-Leu-Gln-CMK) was solved. On the basis of this model and the crystal structure, it was concluded that residues P5 to P3 of the substrate would form an antiparallel β-sheet with segment 164 – 167 of the long strand eII on one side, and with the segment linking domains II and III (residues from 186-191) on the other. Hydrogen-bonding interactions are likely between main-chain N of Thr(P3) – main-chain O of Glu165(M^{pro}), O of Thr(P3) – N of Glu165(M^{pro}), O of Ala(P5) – main-chain N of Gly167 of the protein, as well as between the main-chain N of Ser (P4) and O of Ser189 (protein) (see Fig. 3.35). The fact that the loop connecting domains II and III (residues 183 – 198) is part of this binding site is consistent with the dramatic loss of proteolytic activity upon deletion of domain III (Lu & Denison, 1997; Ziebuhr *et al.*, 1997b; Ng & Liu, 2000). In fact, deletion of the whole polypeptide chain beyond residue 183 leads to total inactivation, whereas

measurable residual activity (0.4%) is observed with a proteinase fragment comprising residues 1 – 199, supporting the important role for the domain II – domain III connecting loop for substrate binding (Anand *et al*., 2002a) (Fig. 3.36B & 3.36D). The feature of substrate binding explains the strong preference for Leu or Ile in the P2 position. The cavity for P2 site is big and hydrophobic enough to accommodate Leu side chains.

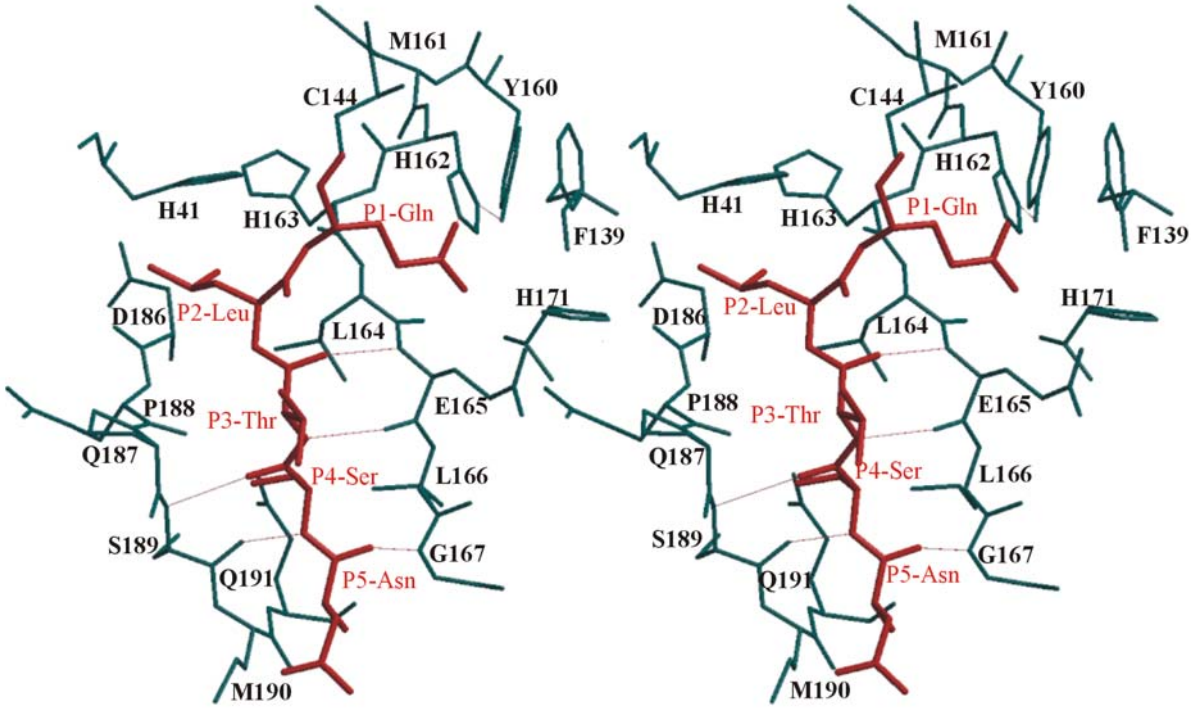


Fig. 3.35 Stereo diagram of a P5–P1 pentapeptide (Asn-Ser-Thr-Leu-Gln, red, corresponding to the TGEV M^{pro} N-terminal autoprocessing site) modeled into the active-site cleft of TGEV M^{pro}. Thin lines depict hydrogen bonds.

3.4.5 TGEV M^{pro} in complex with a substrate-analog chloromethylketone inhibitor

Crystals of the TGEV M^{pro} in complex with peptide (Val-Asn-Ser-Thr-Leu-Gln) were isomorphous to those of free TGEV M^{pro}. The inhibitor was soaked into the crystals, over night (see Section 2.4). They therefore belong to monoclinic space group $P2_1$ with cell dimensions, a = 72.39 Å, b = 158.55 Å, c = 88.20 Å, β = 94.4° and diffract to 2.2 Å resolution. However, the diffraction data were of sufficient quality only to a Bragg spacing of 2.37 Å. Statistics of the diffraction data are given in Table 6.1A (Appendix) and the results of the refinement are summarized in Table 6.1B (Appendix) (Anand *et al*., submitted). Initial electron density maps showed the peptide in the active site of monomer B of the TGEV M^{pro} (Fig. 3.36A), with a covalent bond between the S^{γ} atom of the Cys144 residue and the methylene group of the chloromethyl ketone (Fig. 3.36B). There were small difference

densities in four other monomers (A, C, D, and E) and relatively large density in the F monomer, but not enough to build the whole peptide into it at this stage. Many cycles of manual rebuilding and computational refinement resulted in a final model consisting of residues A1-300, B1-300, C1-300, D1-301, E1-299, and F1-299 residues of TGEV M^{pro} plus 12 residues of peptide in monomers B and F (six residues each). Four MPD molecules fit well near the active sites of monomers A, C, D, E, where there was clearly not sufficient electron density for the peptidic inhibitor. In the initial cycles of refinement, the MPD molecules were removed from the model of the free enzyme to make space for the substrate to be accommodated. It is still unclear why the peptide bound to only two M^{pro} molecules out of six. One reason could be that MPD, which is used in the crystallization solution, competes for the binding in all the molecules but could not succeed in monomers B and F. The other reason could be insufficient amounts of soaking time or inhibitor concentration. The substrate binds at the surface, with only the side chain of Leu (P2) partially buried, whereas of the proteinase, Leu 164 and Leu 166 are buried (Fig. 3.36C). The hexapeptide Z-Val-Asn-Ser-Thr-Leu-Gln-CMK is bound to the S6→S1 specificity subsites of the proteinase. The interactions between substrate and the enzyme are very similar to the previous computer model (see above Fig. 3.35). There are hydrogen-bonding interactions between the main-chain N of Thr(P3) and

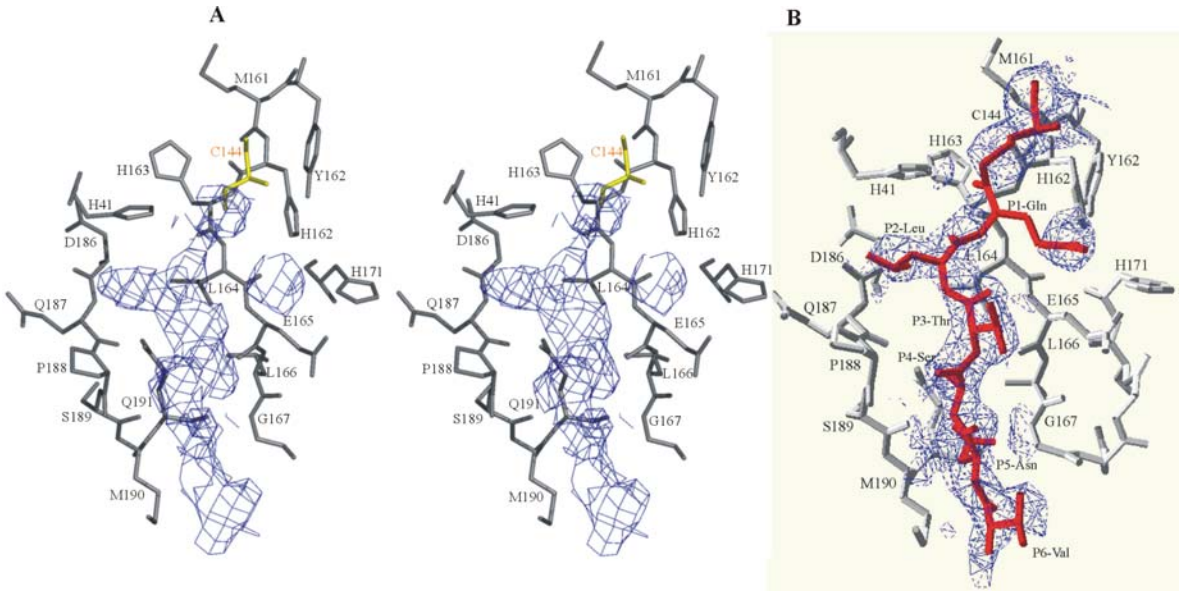


Fig. 3.36**A.** Difference density ($|F_o|$-$|F_c|$) at 2.7 σ above the mean for monomer B of TGEV M^{pro} was clear enough to build the peptidyl inhibitor into it. Similar density was found in monomer F as well. **B.** Shows the refined model of the covalently bound inhibitor, to the SG atom of Cys144, in the electron density. Figure prepared using program Swiss-PdbViewer (Guex *et al*, 1999).

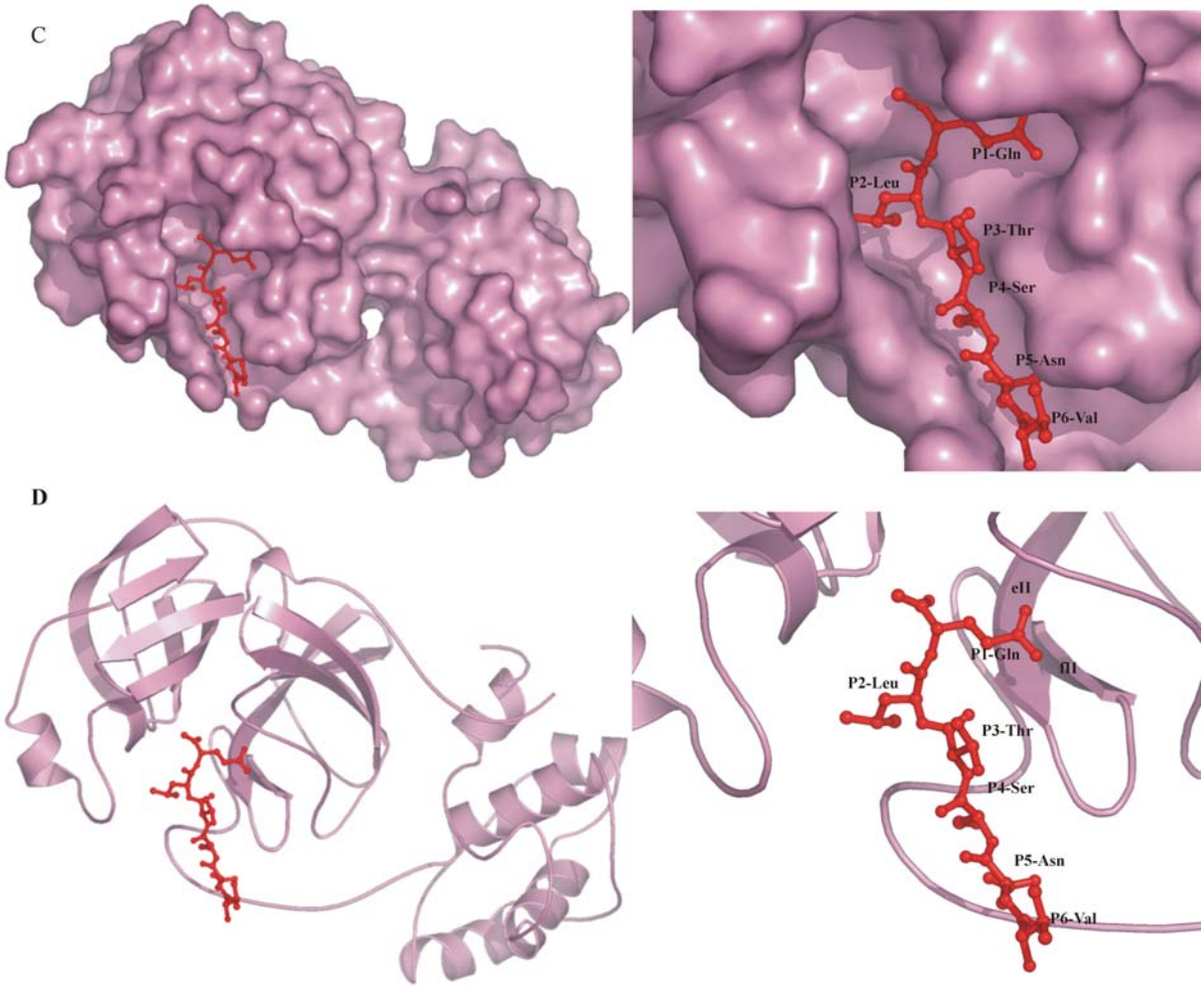


Fig 3.36C Peptidyl chloromethyl ketone inhibitor bound to monomer B of the TGEV M^{pro}. The left panel shows the whole monomer and the right-side panel provides an enlarged view into the substrate-binding site. The peptidic inhibitor is shown in red. D. Cartoon diagram in similar orientation as above, showing the involvement of the long loop connecting domains II and III in positioning the substrate. One of the possible roles of domain I
II is to fix this loop. Figure prepared using program PYMOL (DeLano, 2000).

the main-chain O of Glu165(M^{pro}) 2.90 (± 0.08) Å, O of Thr(P3) and N of Glu165 M^{pro} is 2.94 (± 0.19)) Å, O of Ala(P5) and the main-chain N of Gly167 of the M^{pro} main-chain N of Ser(P4) - main-chain O of Met189 is 2.67 (± 0.35) Å (average over model and structure, average over the monomers is given below). The hydrogen-bonding interactions between main-chain N of Thr(P3) – main-chain O of Glu165(M^{pro}) is 2.93 (± 0.01) Å, O of Thr(P3) – N of Glu165 M^{pro} 2.94 (± 0.05) Å, main-chain N of Ser(P4) - main-chain O of Met189 2.68 (± 0.01) Å (average over monomer B and F).

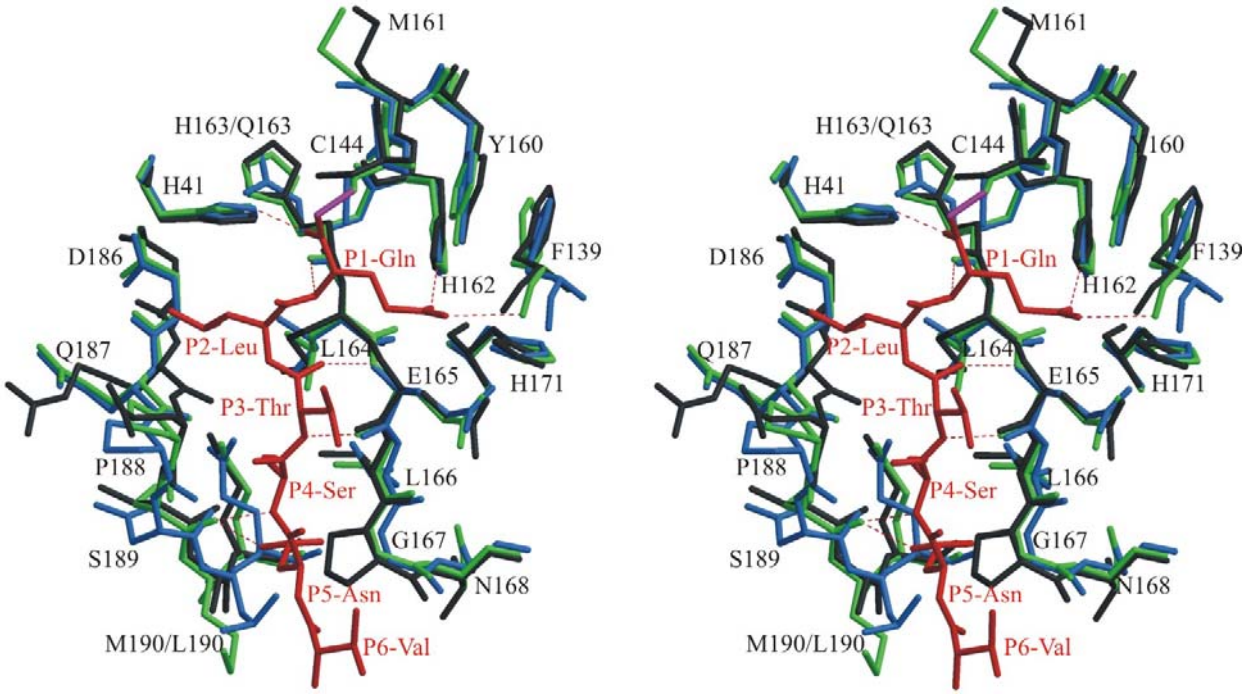


Fig 3.37A. Stereo diagram of a TGEV M^{pro} – CMK inhibitor (substrate analog (P6–P1): Val-Asn-Ser-Thr-Leu-Gln, red, corresponding to the TGEV M^{pro} (green) N-terminal autoprocessing site) complex as described crystallographically. Dotted lines depict hydrogen bonds. Phe139 is stacking to His162. The Tyr-Met-His motif is highlighted. Figure drawn with Molscript (Kraulis, 1991). Substrate binding sites of HCoV M^{pro} are in blue.

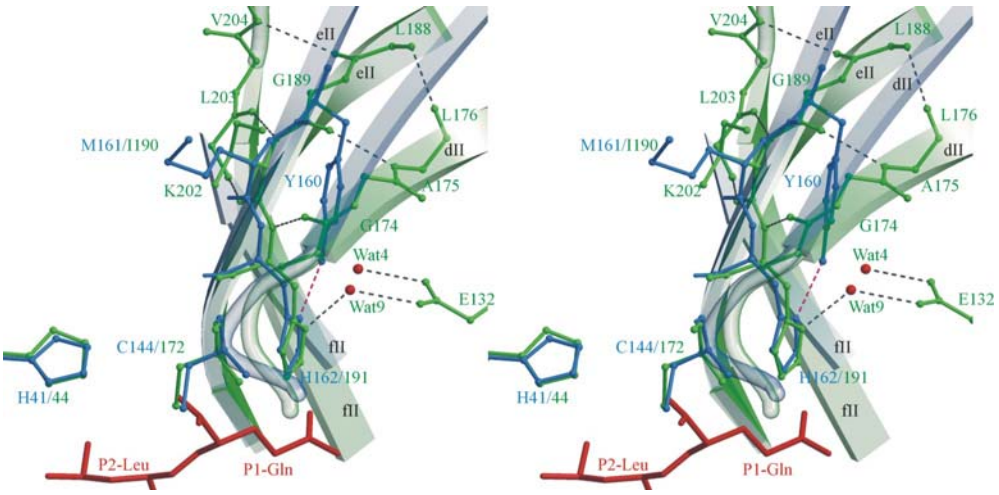


Fig. 3.37B. Stereo diagram of superposition of S1 subsites of TGEV M^{pro} and HAV $3C^{pro}$ showing conserved the YXH and GXH motifs, respectively. In HAV $3C^{pro}$, Gly at this position prevents the disruption of β-sheet (eII-fII); In addition, there is no space for any side chain at this position. Gly189 adopts torsion angles $\varphi = 164^{o}$ $\psi = -165^{o}$ to maintain the β-sheet structure involving strands dII, eII and fII, whereas in TGEV M^{pro} the arrangement of the strands is different and provides more space for the Tyr side chain. P2-P1 positions of CMK inhibitor are shown in red.

3.4.5.1 Substrate specificity

In TGEV and HCoV M^{pro}s, His162 is a critical residue involved in hydrogen bonding to the P1 Gln. The specificity for Gln at the P1 position of the substrate is attributed to the presence of the highly conserved His191 in the S1 pocket in HAV $3C^{pro}$ (Bergmann *et al.*, 1997). In the sequence alignment of M^{pro} with HAV $3C^{pro}$, His 162 corresponds to His191, and the two can be structurally superimposed. In HAV$3C^{pro}$, two water molecules (wat4 & wat9) and the buried Glu132 contribute to the stabilization of the imidazole tautomer of His191 (Fig 3.37B). In TGEV M^{pro}, $N^{\delta 1}$ of His162 and the OH of buried Tyr160 are hydrogen-bonded so that His162 is locked in a neutral single tautomeric form. This further strengthens the role of His162 in substrate binding. In principle, Tyr160 OH can accept a hydrogen bond from His162 $N^{\delta 1}$ or donate one to it, but the OH group strongly prefers to be hydrogen-bond donor. In a database of high-resolution, non-homologous protein structures, 97% of the buried Tyr-OH groups identified donated hydrogen bonds (McDonald & Thornton, 1994). Since His162 $N^{\delta 1}$, is the only nearby hydrogen-bonding partner, it is likely that Tyr160 OH donates a hydrogen bond to $N^{\delta 1}$, thereby stabilizing the His162 tautomer having a proton at $N^{\varepsilon 2}$. This stabilizes the Tyr-X-His (160-161-162) motif (Fig. 3.38), which is an integral region of the substrate-binding pocket and is invariant in all coronavirus main proteinases.

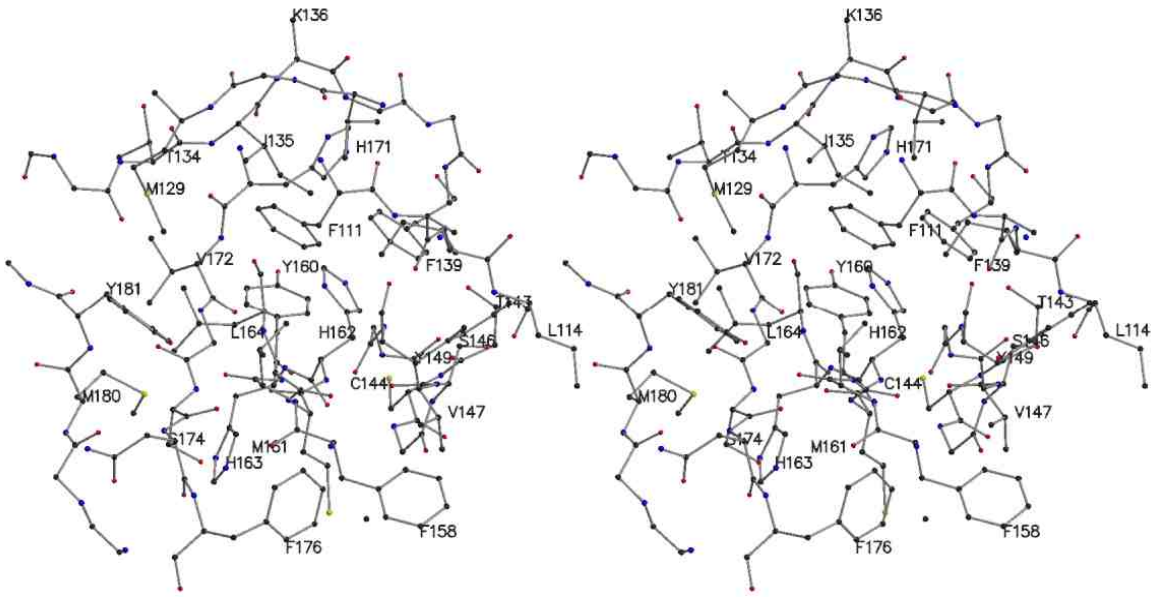


Fig. 3.38 Stereo diagram showing the aromatic cluster around Tyr160 in the TGEV M^{pro} structure. Residues are selected within a radius of 6 $Å^2$ around Tyr160. The later is part of the coronavirus specific Tyr-X-His motif.

In HAV $3C^{pro}$, there is no Tyr needed because of the presence of Glu132, which is neutralizing His191 through two water molecules. In TGEV/HCoV M^{pro}, this role is played by Tyr160. The environment of Tyr160 is rather interesting. It is located at the center of an aromatic cluster built by residues Tyr181, His162, Phe111, His171, Phe139 (Fig. 3.38). The hydrophobic residues Ile105, Ile135, and Val159 make up the remaining environment around the Tyr160 residue. Undoubtedly, mutation of this residue would create a large void that will disrupt the aromatic cluster with significant energetic consequences. Also, His162 forms a bridge between Tyr160 and Cys144; therefore, a mutation would directly affect binding site and disturb the catalytic environment (Fig. 3.31).

The Tyr-X-His motif is a distinct feature of coronavirus proteinases. The important role of histidine162 at this position has been apparent from site-directed mutagenesis studies where the replacement by Ser completely abolished the proteolytic activity (Ziebuhr *et al*., 1997b). This inactivation was selective since a similar replacement of His171, another conserved His residue in this region, was not so detrimental (Ziebuhr *et al*., 1997b). The Tyr-X-His motif has a Met residue in the middle position, which is embedded at the edge of the β-strand eII, which straddles the active-site groove. The Met at the X position is facing away from the aromatic cluster (Fig. 3.38) and not in direct contact with the binding groove. Accordingly, it could be substituted by an Ala residue without significant effects on proteolytic activity (Hegyi *et al*., 2002). The TGEV and HCoV M^{pro} structures clearly show that, despite of its unique sequence context, His162 represents the functional equivalent to the S1 His residue of other viral chymotrypsin-like enzymes.

3.4.5.2 S1 subsite

The S1 subsite of TGEV M^{pro} is formed by residue Phe139, the main-chain atoms of Ile140, Leu164, Glu165 and His171, as well as by the imidazole group of His162. The side chain of Glu165 forms an ion pair (2.96 ± 0.14 Å) with His171. In HCoV M^{pro}, the same residues form the S1 subsite except for a Leu→Ile subsititution at position 164. The ion pair between Gly165 and H171 is preserved in HCoV M^{pro}. This salt bridge is itself on the periphery of the molecule, forming part of the "outer wall" of the S1 subsite. Accordingly, mutants of the HCoV M^{pro}, in which the residue equivalent to His171 had been replaced by Ala, Ser or Thr, retained significant proteolytic activities (Ziebuhr *et al.*, 1997b). Glu165 and His171 correspond to glycines 216 and 226 of chymotrypsin, and to Gly194 and Leu199 of HAV $3C^{pro}$, all of which contribute to the walls of the S1 subsite.

Halfway down the S1 subsite of free TGEV M^{pro}, there is dumb-bell-shaped electron density which we have assigned to two water molecules (W989 and W990), although they are theoretically too close to one another (2.10 ± 0.16 Å). One of them makes a hydrogen bond with $N^{\varepsilon 2}$ of His162, while the second one, unusual for water, makes no additional contacts. In our model of the substrate complex, these two water molecules mark the position of the carboxamide group of the P1 glutamine side chain. This arrangement has been confirmed by the crystal structure of the peptidyl-chloromethyl ketone inhibitor complex of TGEV M^{pro}. A similar arrangement of water molecules is seen in the HCoV M^{pro} crystal structure.

3.4.5.3 Subsites S2 to S4

The hydrophobic S2 pocket is lined by residues Leu164 (side-chain, main-chain contributes to S1; HAV $3C^{pro}$: Ala193; chymotrypsin: Trp215), Pro188 (HAV: Ile216 chymotrypsin: Asn245), His41 (HAV: His44; chymotrypsin: His57), and Thr47 (not present in HAV or CHT). This subsite is relatively large and can easily accommodate the leucine side chain, which is almost invariably present at the P2 position of the substrates. In HAV $3C^{pro}$, the corresponding subsite is formed by different parts of the polypeptide chain. It is also smaller and can accommodate the side chains of serine and threonine (Bergmann *et al*., 1997). The S2 subsite of PV $3C^{pro}$ is a slight depression adjacent to the His40-Glu71 couple and is formed by Gly163, Leu127, Gly128, Gly129 and possibly Arg130 (Mosimann *et al*., 1997).

There is no specificity for any particular side chain at the P3 position of coronavirus M^{pro} cleavage sites (Ziebuhr *et al*., 2000). This agrees with the structure of peptidyl-CMK inhibitor complex in which the P3 residue of the model substrate is oriented towards bulk solvent.

At the P4 position, there has to be a small amino acid residue such as Ser, Thr, Val or Pro because of the congested cavity formed by residues Leu164 (HAV Ala193, CHT Gly215), Leu166 (HAV Gly195; CHT Ser 217), Ser189 main-chain (not present in HAV or CHT), and Gln191 (not present in HAV or CHT).

3.4.5.4 S1' subsite

The TGEV/HCoV M^{pro} substrates have small residues such as Ala and Ser at the P1' position. Only in one TGEV replicase polyprotein cleavage site does Asn occupy the P1' position and competition cleavage experiments have shown that this site is less efficiently cleaved than other sites, indicating the exceptional nature of this substrate (Hegyi & Ziebuhr, 2002). The S1' subsite of the TGEV/HCoV M^{pro} is formed by Leu27, His41, and Thr47, with the latter two residues also being involved in the S2 subsite. Asn and Cys are uncommon as P1' residues outside of the coronaviruses, although there is Asn at the P1' position in rhinoviruses (Blom *et al*., 1996).

3.4.6 Interaction with viral RNA

It has been shown that picornavirus 3C proteinases bind to the 5' NTR cloverleaf structure of the viral RNA (Andino *et al.*, 1993; Leong *et al.*, 1993; Xiang *et al.*, 1995). This interaction is mediated through a conserved KFRDI sequence motif (residues 95-99 in HAV $3C^{pro}$), which is located between domains I and II, as well as small helices and reverse turns (Matthews *et al.*, 1994; Bergmann *et al.*, 1997; Mosimann *et al.*, 1997). The same motif is found in HAV,

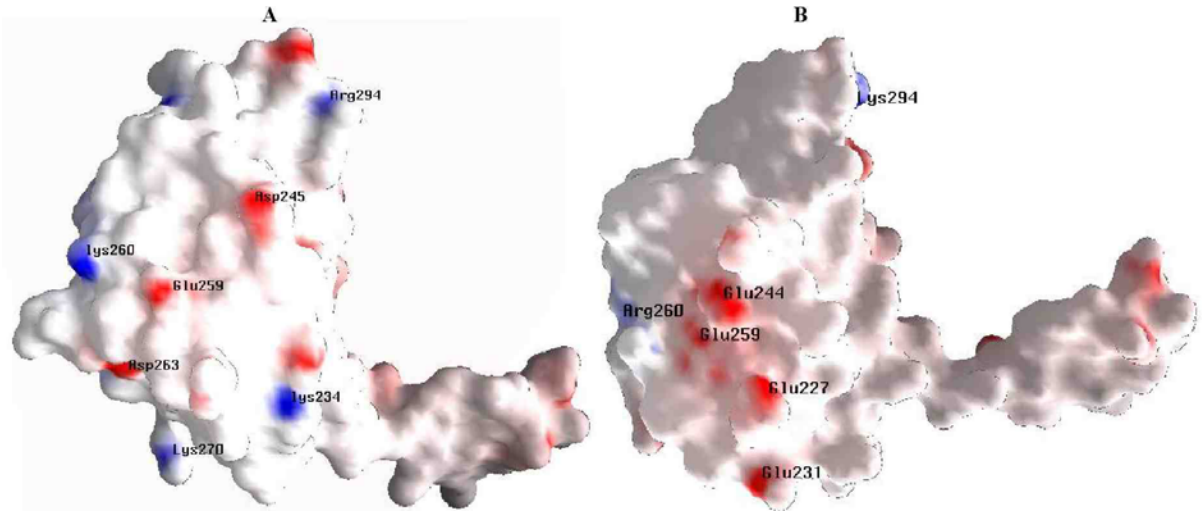


Fig. 3.39 Electrostatic potential over domain III including the long loop (16 residues) connecting domains II and III in **A.** TGEV M^{pro}, **B.** HCoV M^{pro}

poliovirus and rhinovirus proteinases (residues 95 to 99 in HAV, and residues 82 to 86 in poliovirus and rhinovirus). Mutational studies on poliovirus 3C initially implicated these residues as having a function in the picornavirus life cycle, distinct from the proteolytic

activity (Hämmerle *et al.*, 1992). Recognition of the nontranslated regions of the viral RNA genome (5' and 3' NTR) by the 3C subunit of a protein complex appears to be important for the initiation of RNA replication in picornaviruses, but experimental evidence for HAV is not available. One face each of the N- and C-terminal helices (helix A and H), and several of the small helices and reverse turns that connect strands β-bI and β-cI, β-dI and β-eII, β-dII and β-eII form the RNA recognition site of HAV 3C.

Although the $3C^{pro}$ structure is primarily built of β-sheets, helical regions form most of the RNA recognition site. The picornavirus $3C^{pro}$ RNA recognition site is on the side of the molecule opposite to the proteolytic active site. The two-domain β-barrel core structure of HAV 3C is rather rigid and does not allow for conformational changes. Sequence as well as structural alignment shows that the RNA-binding motif (KFRDI) is not conserved in coronavirus proteinases. In HCoV M^{pro} and TGEV M^{pro}, the residues corresponding to positions to KFRDI (95-99) are VNTPE (93-94) and TNTPR (97-99), respectively.

The α-helical C-terminal domain III could in principle, be a candidate for RNA recognition. The domain was surveyed in both HCoV and TGEV M^{pro} for positive surface electrostatic potential to identify potential RNA-binding sites (Fig. 3.39). Since there is no overall basic character nor distinct patches of basic or aromatic residues, a role in RNA binding seems not probable for the extra C-terminal domain of the M^{pro}s. In summary, no RNA-binding motif is present in coronavirus main proteinases and all of the structural elements forming the RNA recognition site in picornavirus 3C proteinases are either missing or very different in the M^{pro} structures. Therefore, a role of the coronavirus main proteinases in RNA binding is unlikely.

3.4.7 Chain termini and autoprocessing

Autoproteolysis is one of the mechanisms RNA viruses have evolved to regulate viral gene expression. Polyproteins containing protein domains with distinct functions are separated co-translationally or post-translationally by autoproteolytic processing. The TGEV/HCoV M^{pro} crystal structures provide insight into the molecular details of these processes.

The following observations for monomer A of TGEV M^{pro} hold true for all other monomers. In the dimer, the N-terminal segment 1 – 8 of monomer A is squeezed in between domains II and III of the same monomer and domains II and III of monomer B (Fig. 3.40A, B

& 3.41). Residues 6A to 8A form a short β-strand interacting with strand cII of monomer B (at Val124B). An intermolecular ionic interaction between NH2 of Arg4A and OE2 of Glu286B (5.0 ± 1.55 Å) and NH1 of Arg4A and OE1 of Glu286B (5.93 ± 1.02 Å) appears to play a role in positioning the N-terminal residues. Because of the twofold non-crystallographic symmetry (NCS), the same interaction occurs between Arg4B and Glu286A. In addition, the side-chain amino group of Lys5A makes strong intramolecular hydrogen bonds with Ser110A O^{γ} of domain II (2.83 ± 0.15 Å), and with the Glu286A main-chain oxygen (2.80 ± 0.07 Å) as well as with Glu291A $O^{\varepsilon 1}$ (2.74 ± 0.13 Å) of domain III. The residues N-terminal to this pair of positively charged amino acids (Arg4, Lys5) also interact strongly with both proteinase monomers. The side chain of Leu3A completes a hydrophobic patch on domain III that includes Phe206A, Ala209A, Phe287A, Val292A, Gln295A, and Met296A; these residues belong to helices B and E. All members of the coronavirus proteinase family have a hydrophobic residue in position 3, and an absolutely conserved glycine in position 2. The latter residue adopts the α_L conformation, which is easily accessible only to glycine. This conformation ensures that the N-terminal segment fits in between domains II and III of the same monomer and domain II of the second monomer in the dimer.

Even more interestingly, Ser1A is in contact with residues participating in the substrate-binding site of monomer B. Its main-chain NH_3^+ group makes a salt bridge (4.30 ± 1.22 Å) to the carboxylate of Glu165B. This glutamate, which is absolutely conserved among coronaviruses, is part of the S1 subsite (see above), where it also interacts with His171. Although these two side chains form the 'wall' of the specificity site, they have their polar groups oriented towards the surface of the proteinase molecule and away from the P1 glutamine of the substrate (Fig. 3.36D, 3.37A). Most of the interactions between the N-terminus of molecule A and the region next to the S1 subsite of molecule B constitute a perfect fit. This suggests that the tight dimer seen in the crystal is stable in solution as well as under physiological conditions. This is also supported by the fact that the same situation is observed in the crystal structure of the HCoV M^{pro}, which also crystallizes as a dimer, but under different conditions and in a different unit cell (Section 3.2.5.3, Table 3.6). Furthermore, most of the residues stabilizing the inter-domain contacts in the TGEV M^{pro} are conserved among coronavirus M^{pro} enzymes.

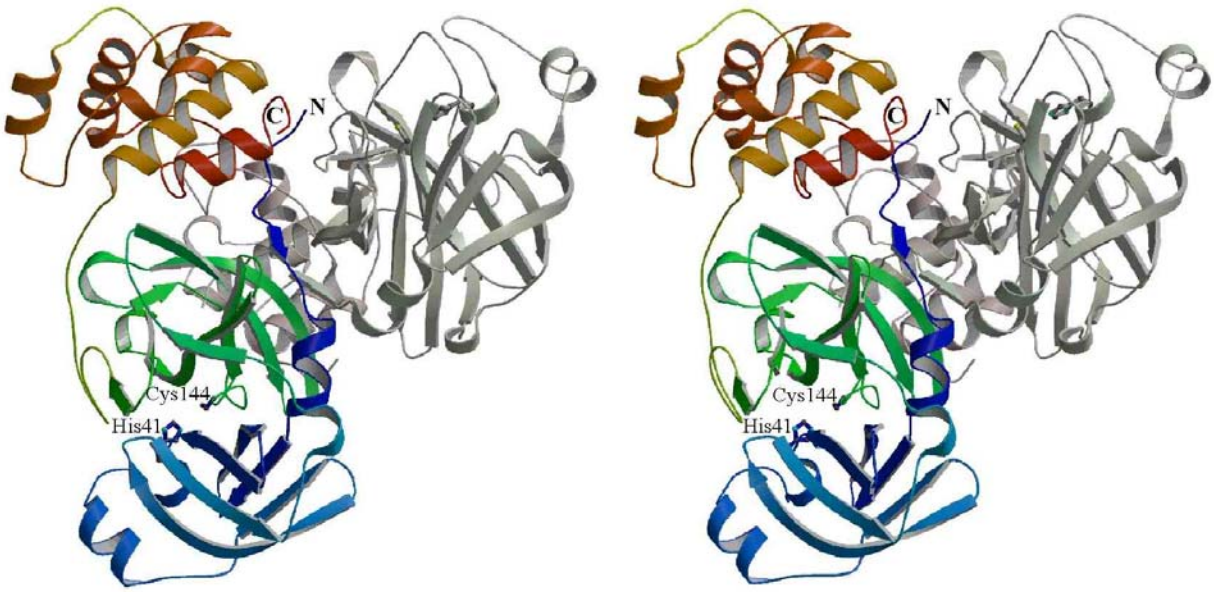


Fig. 3.40 **A.** Stereo figure showing the N-terminal segment located between the two monomers A and B. Molecule A is colored from blue at the N-terminus, via green (domain II), to red (C-terminus), while molecule B is shown in grey. The catalytic Cys144 and His41 residues are shown in ball-and-stick in both monomers. The N- and C-termini of molecule A are indicated.

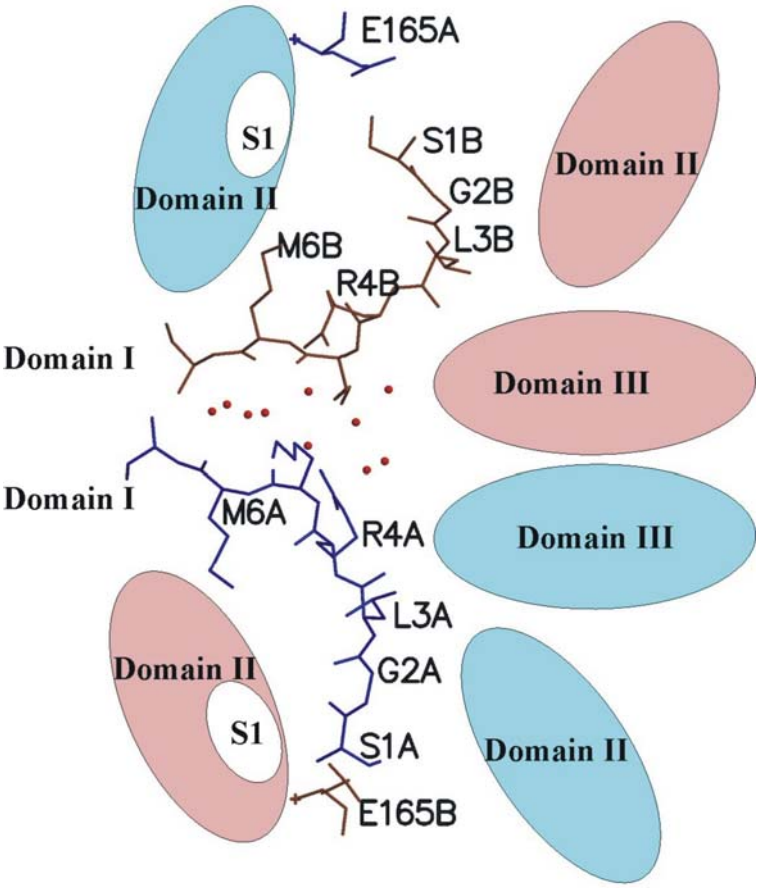


Fig. 3.40 **B.** Schematic representation of the inter- and intramolecular contacts made by the N-termini of molecules A and B of the TGEV M^{pro} dimer. Amino acid residues and domains of molecule A and B are shown in blue and brown, respectively. The S1 binding sites, which are part of domain II, are indicated and the water molecules (red dots) filling the space between the N-termini of molecules A and B are depicted by red spheres.

An important role for the intricate interactions made by the N-terminal residues (Fig. 3.41) is supported by the fact that the peptidolytic activity of TGEV M^{pro} drops to 0.3 – 0.6% after deletion of residues 1 – 5 of the polypeptide chain (Anand *et al*., 2002a). This observation suggests that the N-terminal residues are responsible for fixing the mutual orientation of domains II (through Lys5, Met6) and III (through Leu3, Arg4, Lys5) of its own protomer, thereby ensuring that the loop (residue 184-199) connecting the two domains and involved in binding of substrate residues P5 to P3 (see above) has the proper orientation. In addition, they are also involved in orienting domains II and III of the other protomer, through intermolecular interactions involving the amino terminal NH_3^+ group and Arg4, respectively.

Providing a specific binding site for the N-terminal peptides may have yet another important function. Ser1A, which is the P1' residue of the autocleavage reaction of TGEV M^{pro}, is 11.9 ± 1.6 Å from the active site Cys144B S^{γ} of the second molecule in the dimer and approximately 13.8 Å from the N-terminal nitrogen of the P1' residue which was modeled into the substrate-binding site. In contrast, the active site cysteine of molecule A is as much as 34.2 ± 0.9 Å away from its own N-terminus. Given the fact that the P' residues in serine and cysteine proteinases constitute the leaving group of the cleavage reaction and are usually not subject to stringent specificity requirements in coronavirus M^{pro}s, it is quite conceivable that, after autoproteolysis, the N-terminus of one monomer slides over the active site of the partner monomer and adopts the position seen in the M^{pro} crystal structures, *i.e.,* with Ser1A interacting with Glu165B at the 'outer wall' of the S1 subsite. This, in turn, would suggest that the dimer corresponds to the product of the autolysis reaction and that this occurs in *trans*. This interpretation is also consistent with previously published data from the MHV system. For example, immuno-precipitation experiments revealed that, initially, the MHV M^{pro} is part of a larger, 150-kDa precursor protein in which the proteolytic domain is flanked by hydrophobic domains (Schiller *et al*., 1998). Genetic data indicate that these precursor molecules are functional (Siddell *et al*., 2001) and one of these functions could be to direct the polyprotein precursors to intracellular membranes, *i.e*. the site of replicase activity. The observed delay in the release of the mature, 27-kDa MHV M^{pro} would argue against co-translational processing in *cis* and favor the idea that the MHV p150 concentration required for efficient intermolecular M^{pro} autoprocessing is not reached early in infection. Thus, components of the replication complex could be anchored to membranes in an uncleaved form and only later, when the precursor proteins accumulate to high local concentrations (Schiller *et al*., 1998; van der Meer *et al*., 1999), will M^{pro} dimerize and release itself by

intermolecular cleavage. This would then trigger the complete spectrum of *trans*-processing reactions in order to activate the preformed replication/transcription complex or to modulate its functions.

3.4.8 Role of Domain III

To corroborate the hypothesis of an involvement of domain III in proteolytic activity, an additional set of M^{pro} mutants was characterized by John Ziebuhr, in which the structural information was used to *completely* remove domain III. In these experiments, the danger of domain III misfolding, which might have been the cause of M^{pro} inactivation in previous studies using randomly "truncated" coronavirus main proteinases (Lu and Denison, 1997; Ziebuhr *et al.*, 1997b; Ng and Liu, 2000), was removed. The TGEV M^{pro} deletion mutants tested by Ziebuhr for activity comprised (i) domains I and II ($M^{pro}\Delta184–302$), (ii) domains I and II together with the entire loop region connecting domain II and III ($M^{pro}\Delta200–302$) or

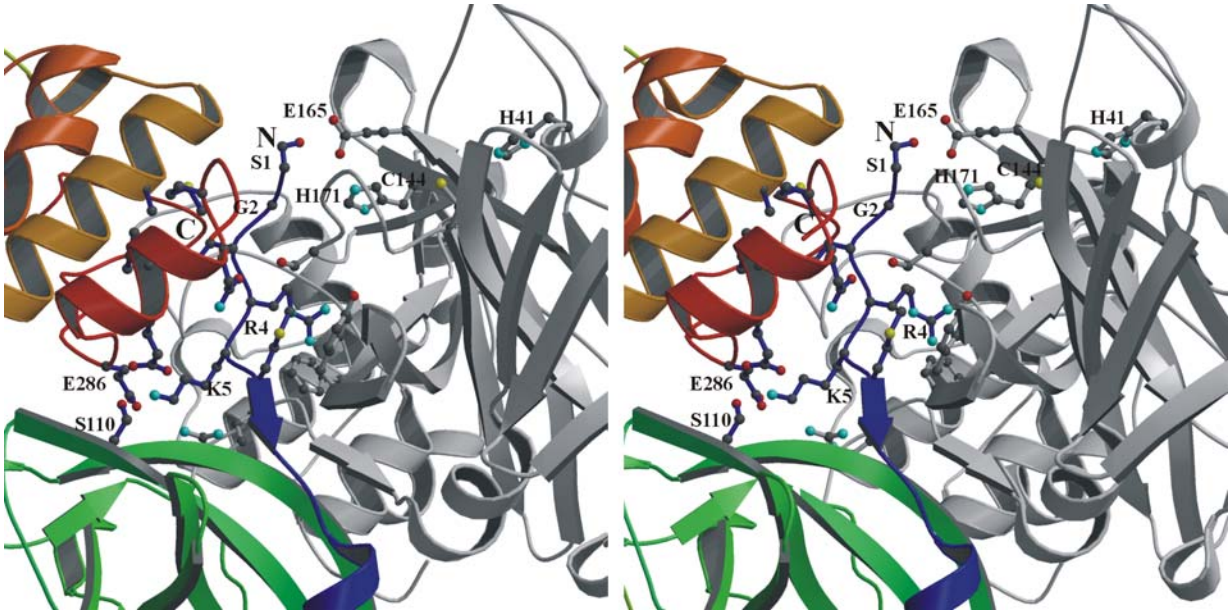


Fig. 3.41 Stereo figure showing the intra- and intermolecular contacts of the TGEV M^{pro} N-terminus. Detailed view of the interactions made by the N-terminal segment (blue) and domain II/III of monomer A as well as domain II/III of monomer B. Residues critically involved in these interactions are designated by the single-letter code and shown in ball-and-stick representation. The N- and C-termini of molecule A are indicated.

(iii) Domains I and II combined with the loop region but lacking the five N-terminal residues (M^{pro}Δ1–5/Δ200–302). As Table 3.12 shows, M^{pro}Δ200–302 had clearly detectable (albeit significantly reduced) activity (0.4% of M^{pro}). Similarly, the mutant M^{pro}Δ1–5/Δ200–302 had significantly reduced activity (0.6% of M^{pro}). In sharp contrast, no activities at all were detectable for M^{pro}Δ184–302 and the active-site mutant, M^{pro}-C144A (the latter being used as a negative control) (Anand et al., 2002a).

Table 3.12 Enzymatic activities of TGEV M^{pro} mutants (Ziebuhr, personal communication)

Plasmid	Oligonucleotides used for cloning or mutagenesis (5' → 3')	Protein	M^{pro} amino acids	Activity (%)[a]
pMal-M^{pro}	TCAGGTTTGCGGAAAATGGCAC, AAAAGGATCCTTACTGAAGATTTACACCATACATTTG	M^{pro}	Ser1 – Gln302	100
pMal-M^{pro}Δ184-302	TCAGGTTTGCGGAAAATGGCAC, AAAGGATCCTTAACCACCGTACATTTCTCCTTCAAAATT	M^{pro}Δ184-302	Ser1 – Gly183	<0.02
pMal-M^{pro}Δ200-302	TCAGGTTTGCGGAAAATGGCAC, AAAGGATCCTTATGACATGACATTAGTACCTTCCAATTG	M^{pro}Δ200-302	Ser1 – Ser199	0.4
pMal-M^{pro}Δ1-5/Δ200-302	ATGGCACAGCCTAGTGGTCTTGTA, AAAGGATCCTTATGACATGACATTAGTACCTTCCAATTG	M^{pro}Δ1-5/Δ200-302	Met6 – Ser199	0.6
pMal-M^{pro}Δ1-5	ATGGCACAGCCTAGTGGTCTTGTA, AAAAGGATCCTTACTGAAGATTTACACCATACATTTG	M^{pro}Δ1-5	Met6 – Gln302	0.3
pMal-M^{pro}-H163L	GTATACATGCATCTCTTAGAACTTGGAAATGGCTCGCAT, TCCAAGTTCTAAGAGATGCATGTATACAAAATAGAGAAT	M^{proo}-H163L	Ser1 – Gln302 (His163 → Leu)	98
pMal-M^{pro}-C144A	AGCTGGTACTGCTGGATCAGTAGGTTATGTGTTAGAA, CTACTGATCCAGCAGTACCAGCTATAAAAGATCCTTT	M^{pro}-C144A	Ser1 – Gln302 (Cys144 → Ala)	<0.02

[a] Proteolytic activities were determined using a peptide-based cleavage assay (Ziebuhr *et al.*, 1997). The sequence of the 15-mer substrate peptide, H_2N-VSVNSTL**QS**GLRKMA-COOH was derived from the N-terminal M^{pro} autoprocessing site (residues shown in bold-face indicate the scissile bond). The activity of wild-type M^{pro} (encompassing 302 residues) was taken as 100% and the mean value of three experiments, which did not vary by more than 15%, is shown (Anand et al., 2002a).

The fact that residues 184–199 proved to be indispensable for proteolytic activity is in agreement with the structure of the peptidyl chloromethyl ketone inhibitor complex of TGEV M^{pro} (Fig. 3.36, 3.37) in which residues of the loop are predicted to be critically involved in the formation of a β-sheet-type structure with the substrate (Section 3.4.4). The data also show that an intact N-terminus and the C-terminal domain are required for full activity. The

structure suggests that the additional α-helical domain III as well as the N-terminal residues help fix domains II and the loop 184-199 in a catalytically competent orientation.

It will be interesting to investigate whether similar mechanisms are also operating in other 3C-like proteinases with (smaller) C-terminal domains (*e.g.*, arteriviruses and potyviruses; Ziebuhr *et al.*, 2000; Hegyi *et al.*, 2002).

3.5 Relationship of coronavirus M^{pro} with viral and cellular homologs

The TGEV/HCoV M^{pro} is unique as it represents one of the largest RNA virus proteinases (~30 kDa). Usually chymotrypsin-like proteinases from other viruses have only around 180 residues. The size difference is due to the extra C-terminal region of approximately 100 amino acids. Coronavirus M^{pro} displays a very low overall sequence similarity (<20%) to other proteases. Basically, the similarity is limited to the regions of active-site residues.

The TGEV/HCoV proteinase structure represents a new class of cysteine proteinases which combine the common two β-barrel serine proteinase like fold with an additional C-terminal domain as previously predicted (Bazan & Fletterick, 1988; Gorbalenya *et al.*, 1989) (Fig. 3.13 & 3.43). In fact, a novel structural fold (the C-terminal domain) may be employed in this particular case to control the proteolytic activity, even though the mechanism of control cannot easily be derived from its structure. It is found that deletion mutants (Ser1-Gly183 and Ser1-Ser199) removing all of the C-terminal residues are more than 200 fold less active in the standard peptide-cleavage assay (see above).

A three-dimensional comparison of domains I and II of TGEV/HCoV proteinase reveals a significant relationship with the main-chain folds of the picornavirus HAV 3C and chymotrypsin-like serine proteinases. The three-dimensional structural comparison among all three (excluding HCoV M^{pro}) gave an rms deviation of ~1.8 Å for approximately 100 out of 200 residues (Fig. 3.42). Even though the active-site histidine and cysteine residues superimpose, and a two β-barrel fold is present in both proteins, the differences are quite evident. The bII and cII strands are extremely shortened and there is different connectivity of dII and eII in HAV 3C. Additionally, the C-terminus folds back in chymotrypsin and HAV $3C^{pro}$ to get in contact with domain I while there is no such interaction in TGEV and HCoV proteinases.

Chymotrypsin, trypsin and elastase have very similar three-dimensional structures but different specificity. They cleave adjacent to bulky aromatic side chains, positively charged side chains, and small-uncharged side chains, respectively. In trypsin and chymotrypsin, a glycine residue allows the side chain of the substrate to penetrate into the interior of the specificity pocket. In elastase, Val and Thr fill up most of the pocket, so that proteinase cleaves adjacent to small-uncharged side chains.

The substrate specificity of TGEV/HCoV proteinase resembles that of many other 3C/3C-like proteinases (Kräusslich & Wimmer, 1988; Dougherty & Semler, 1993; Blom *et al*., 1996) in so far as the P1 position of the substrate is exclusively occupied by Gln and small residues (Ser, Ala, Asn, Gly and Cys) are found at the P1' position. The P2 and P4 positions are mostly conserved, with bulky hydrophobic residues (mainly Leu) at P2 and Val, Thr, Ser (and Pro) at P4 being clearly favored. However, their three-dimensional structure or sequence is quite different from any picornavirus 3C proteinase.

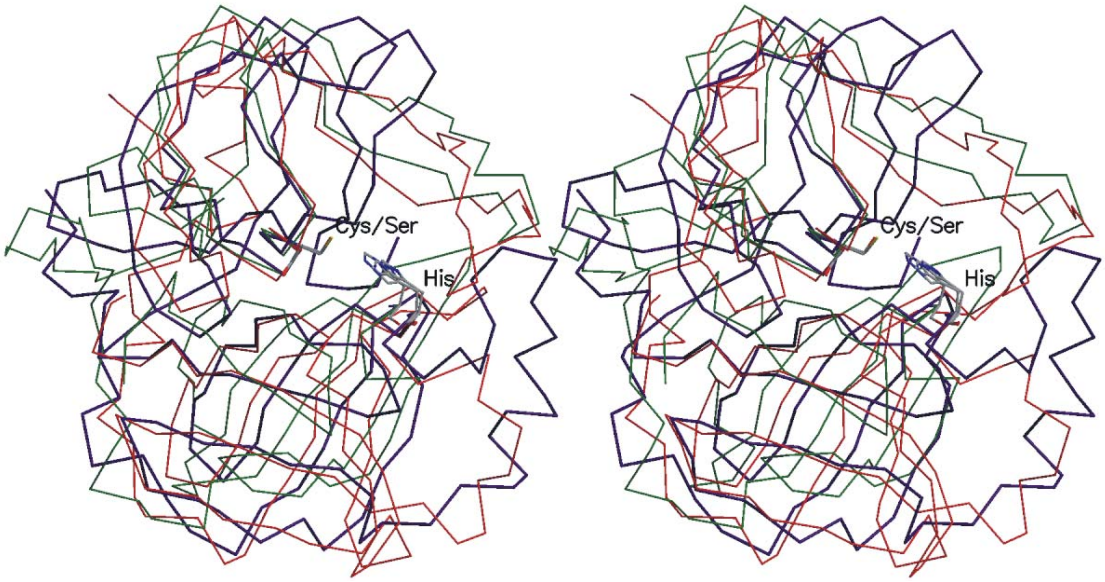


Fig. 3.42 A stereo diagram of the C^{α} trace of the least-squares superposition of TGEV M^{pro} (thick line, violet), HAV-3C (red) and chymotrypsin (green). The Cys/Ser nucleophile and the His general acid/base catalyst at the active site of the three enzymes are also shown. The root-mean square deviation is 1.8 Å for 100 equivalent C^{α} atoms that superpose within 3.8 Å. The active sites superpose closely whereas other regions including surface loops and turns deviate more.

In 3C and 3C-like proteinases, Cys replaces the nucleophilic Ser and, in a subset of viruses, Glu replaces the Asp of the catalytic triad found in cellular proteinases (Bazan & Fletterick, 1988; Gobalenya *et al*., 1989a; Matthews *et al.*, 1994). The coronavirus M^{pro} seems to lack a conserved acidic residue that would be equivalent to the catalytic Asp (Glu) of 3C proteinases

In thiol proteinases such as papain, the sulphydryl group of the active site cysteine shows high nucleophilicity – the sulfur atom forms a thiolate anion/imidazolium couple with His159 at neutral pH. The oxygen of the Asn175 side-chain is hydrogen bonded to His159 forming an Asn175-His159-Cys25 triad reminiscent of to the Asp-His-Ser triad in chymotrypsin as described above (Section 3.4.1). The Asn175-His159 hydrogen bond is approximately collinear with the Cγ-Cβ bond of His159, allowing some rotation of the imidazole ring about this axis without disrupting the bond. Indeed, the function of Asn175 may be modulation of the rotation of the imidazole ring during catalysis. The electrostatic contribution of the oxyanion pocket in cysteine proteinase catalysis appears to be somewhat less than that in serine proteinase catalysis (Menard *et al.*, 1995). In papain, the oxyanion pocket dipoles are derived from the main-chain amide of Cys25 and the side-chain amide of Gln19.

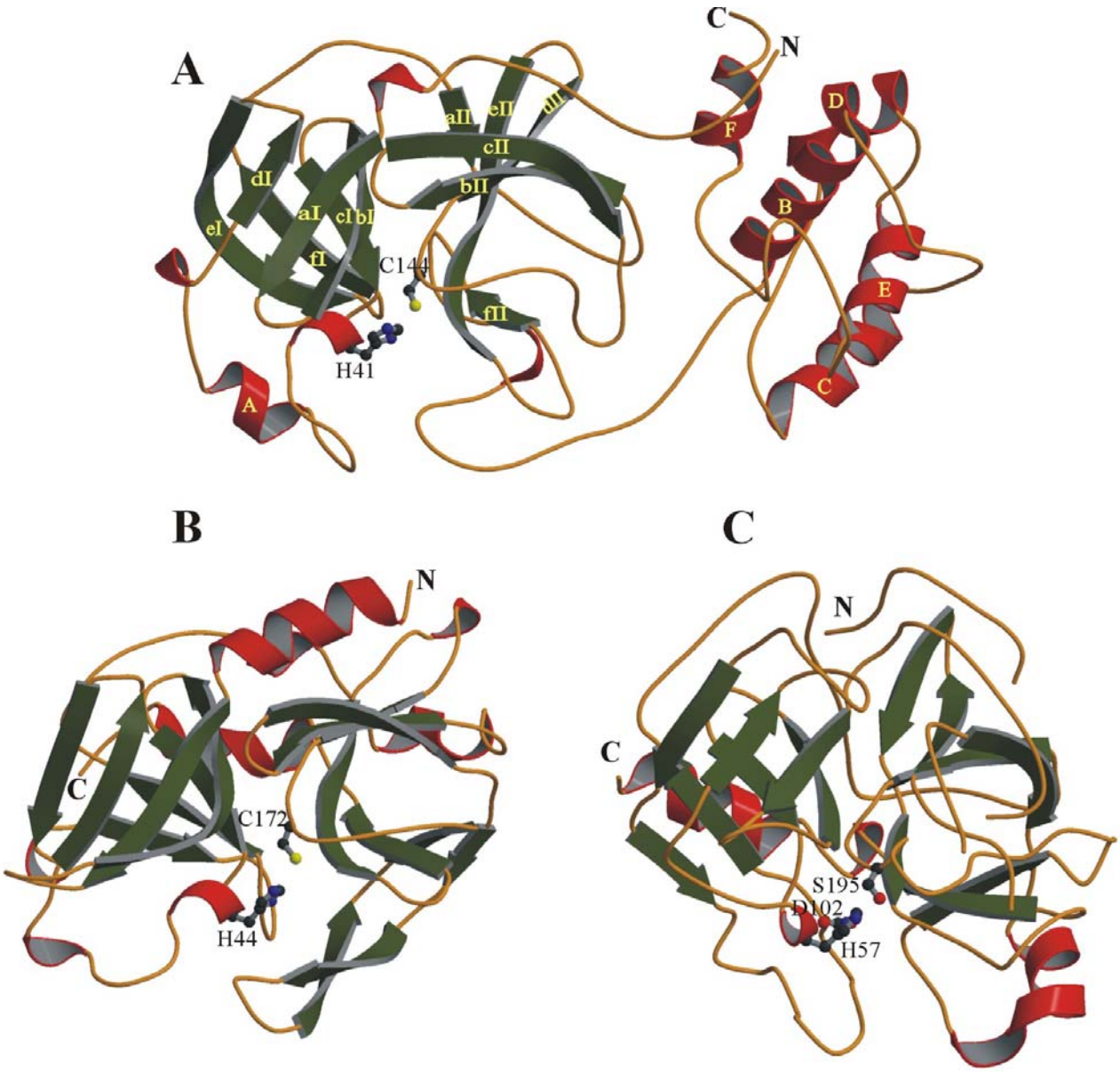


Fig. 3.43 Overall fold of TGEV M^{pro} together with the structures of viral and cellular homologs. (A) TGEV M^{pro} with the two β-barrel domains and the α-helical C-terminal domain. β-strands and helices are represented as arrows and cylinders, respectively. The structures of HAV $3C^{pro}$ (PDB code: 1HAV) (B) and α-chymotrypsin (4CHA, residues 12-15 and 147-148 are excised) (C) are shown for comparison.

In chymotrypsin-like proteinases, the oxyanion hole serves to stabilize the negative charge developing on the P1-carbonyl oxygen atom during the formation of the tetrahedral intermediate. The oxyanion hole consists of a type-II reverse turn preceding the nucleophile and has two main-chain amides pointing towards the P1-carbonyl oxygen atom. The structure of the oxyanion hole in the A monomer of TGEV/HCoV proteinase is shown in Figure 3.31A & 3.31B. Thr143 forms a hydrogen bond with Ala140 (main-chain). This hydrogen-bond seems to stabilize the oxyanion hole by anchoring the two consecutive type-II reverse turns formed by residues 140-143 to the hairpin loop between strands bII and cII (residues Thr117-Cys120). The NH of T143 forms an H-bond to I140 C=O. Additionally, the T143$O^{\gamma}H$ form H-bond to I140 C=O. Gly142 C=O receives an H-bond from N128$^{\delta 2}$. The next type II turn is: T143-C144-G145-S146. Similar hydrogen bonds from S146 to C=O of T143 as in the other preceding type II turn are present here. The hydrogen-bonding pattern is similar to that in bacterial chymotrypsin-like proteinases such as α-lytic proteinase (αLP; Fujinaga *et al.*, 1985) and SGPA. These chymotrypsin-like serine proteinases have the sequence Gly193-Asp194-Ser195-Gly196 around the nucleophile, which corresponds to Gly142, Thr/Ala143, Cys144 and G145 of TGEV/HCoV proteinase. The threonine forms a hydrogen bond to the main-chain N at residue -3 from the nucleophile in its own reverse turn in order to stabilize the oxyanion hole (the proximal Tyr117 OH forms an equivalent hydrogen bond in HCoV M^{pro}). This threonine is not present in picornavirus proteinases nor in chymotrypsin. Cys144 is situated in the *i+1* position of the type-II β-turn. His41 is situated in the 3_{10} helix located after β-strand (cI). This β-strand is in a position to form an antiparallel, β-sheet-type hydrogen bonding interaction with a bound substrate.

3.6 CONCLUSIONS

The 60-selenium atom substructure that has been solved for phasing is amongst the largest that has been successfully used in *de novo* MAD structure determination (Deacon and Ealick, 1999). This should encourage the use of selenomethionine as an anomalous scatterer to solve even larger structures. The present work, is the first successful report of the three-dimensional structure of the family *Coronaviridae*.

The three-dimensional structure of TGEV M^{pro} shows that coronaviruses have evolved proteinases in which a thiolate-imidazolium catalytic dyad has been combined with the two-β-barrel fold present in chymotrypsin-like serine proteinases. In agreement with previous studies (Allaire *et al.*, 1994; Bergmann *et al.*, 1997), the M^{pro} structure supports the

hypothesis that the catalytic centers (and mechanisms) used by the eukaryotic prototype cysteine and serine proteinases may only represent variations of a common theme.

The additional α-helical domain III, together with the N-terminal residues, appears to be the involved in proteolytic activity by maintaining the proper positioning of the presumed substrate-binding loop, 184-199. In the model, the three-domain structure ensures the release of the M^{pro} N-terminus by intermolecular autoprocessing, and, at the same time, prevents product inhibition of M^{pro} by providing a proper interaction site for the N-terminal and C-terminal segment. Once bound at this site, residues 1 – 5, in turn, help fix the M^{pro} domains II and III and the loop intervening in a catalytically competent orientation.

The structure also shows that binding of the coronavirus main proteinases to the viral RNA is unlikely, in contrast to the findings for picornavirus 3C proteinases.

Finally, the architecture of the substrate-binding site explains the pronounced substrate specificity of coronavirus main proteinases and provides a basis for the rational design of specific inhibitors that might be used to control and prevent the spread of coronavirus infections.

4. SUMMARY

In this thesis, the ~33kDa structures of porcine transmissible gasteroenteritis virus main proteinse (TGEV M^{pro}) and human coronavirus virus main proteinase (HCoV M^{pro}), the largest RNA virus proteinases known to date, have been determined by X-ray crystallography. The purified proteins were provided by Dr. Ziebuhr (Würzburg). The structure solution was based on a multifarious approach, which included the generation of SeMet-substituted protein crystals and usage of a nonconventional cryprotectant (mustard oil). Data were collected from single crystals held in a stream of nitrogen gas at 100K, using synchrotron beamlines at DESY (Hamburg) and ELETTRA, (Trieste). The phase solution for the TGEV M^{pro} structure was obtained based on MAD data, which were used to locate 60 selenium sites using direct methods (*SnB*). The identification of the 60 selenium sites proved to be a challenging test of the robustness of current phasing strategies for SeMet-based MAD structure determination. Taking TGEV M^{pro} as a model, the structure of HCoV M^{pro} has been solved by the molecular replacement method.

Both TGEV and HCoV M^{pro} form tight dimers; TGEV M^{pro} has three dimers in the asymmetric unit whereas HCoV M^{pro} has only one. All dimers are related by twofold non-crystallographic symmetry. Each monomer has three domains; the first two domains have β-barrel folds similar to those of chymotrypsin-like serine proteases, and the third domain is a unique arrangement of five antiparallel α-helices. Domains I and II contain six β-strands each. In domain I, there is a short helix between strand cI and dI (the HCoV M^{pro} has an additional short N-terminal helix).

The three-dimensional structure of TGEV M^{pro} shows that coronaviruses have evolved proteinases in which a thiolate-imidazolium catalytic *dyad* has been combined with the two-β-barrel fold present in chymotrypsin-like serine proteinases. In agreement with previous studies (Allaire *et al.*, 1994; Bergmann *et al.*, 1997), the M^{pro} structure data strongly support the hypothesis that the catalytic centers (and mechanisms) used by the eukaryotic prototype cysteine and serine proteinases may only represent variations of a common theme. The active sites of both M^{pro}s are identical and the catalytic site residues are made up of Cys144 and His41. In the case of TGEV, Cys144 is oxidized up to sulfonic acid whereas in HCoV M^{pro}, it is oxidized up to the sulfenic acid. The distance (in TGEV M^{pro}) between Cys144 and His41 is larger than the corresponding Cys-His distances in the picornavirus protomers.

To study the active-site geometry in more detail, the structure of TGEV M^{pro} in complex with TLCK has also been solved. The TLCK is found to be covalently bound to the active-site Cys144 but also to Cys120 and Cys284 which are exposed to solvent. Therefore, binding of TLCK did not reveal significant information about the residues involved in the active site. This lead to the design of a sequence-specific substrate-analog peptidyl (P6-P1) inhibitor. The structure of TGEV M^{pro} in complex with this inhibitor was solved by the molecular replacement at 2.37 Å resolution.

The TGEV-CMK inhibitor complex (substrate analog, P6-P1, Val-Asn-Ser-Thr-Leu-Gln) structure shows the complementary nature of the binding surfaces. It reveals that the P4-P1 residues of the peptide assume a common main chain conformation when bound to these proteinases. The residues P5 to P3 form an antiparallel β-sheet with segment 164-167 of the long strand eII on one side, and with the segment 186-191 (which links domain II and III) on the other.

The N-terminal segment plays an important role in fixing the dimer by interacting with domains II and III of its own monomer and domains II and III of the other monomer in the dimer. Most of the interactions are between the N-terminal segment of one monomer with the substrate binding-site residues of the other monomer. This suggests an important role of this stretch of residues in the autoprocessing function of these proteinases. The additional α-helical domain III appears to be the driving force for dimerization of M^{pro}. In this model, the three-domain structure ensures the release of the M^{pro} N-terminus by *inter*molecular autoprocessing and, at the same time, prevents product inhibition of M^{pro} by providing a proper interaction site for the N-terminal segment created after autocleavage. It remains to be investigated whether similar mechanisms are also operating in other 3C-like proteinases with (smaller) C-terminal domains (*e.g.*, arteriviruses and potyviruses; Ziebuhr *et al*., 2000; Hegyi *et al.*, 2002; Barrette-Ng *et al*., 2002). In picornavirus 3C proteinases for example, the refolding of the released N-terminus into a stable helix is assumed to prevent the self-inactivation of $3C^{pro}$ following presumed *intra*molecular N-terminal cleavage (Khan *et al.*, 1999). The structure also shows that binding of the coronavirus main proteinases to the viral RNA is unlikely, in contrast to the findings for picornavirus 3C proteinases.

The current structure analyses of the coronavirus main proteinases provide considerable insight into the structure-function mechanisms of this family of enzymes and also help in understanding the role of M^{pro} in the coronavirus life cycle. The results are instrumental in the structure-based development of lead compounds for anticoronaviral therapy of a variety of coronavirus diseases in humans, domestic animals, and livestock.

5. REFERENCES

Adam PD, Pannu NS, Read RJ & Brünger AT (1997). Cross-validated maximum likelihood enhances crystallographic simmulated annealing refinement. *Proc. Natl. Acad. Sci. USA* **94**, 5018-5023.

Allaire M, Chernaia MM, Malcolm BA & James MNG (1994). Picornaviral 3C cysteine proteinases have a fold similar to chymotrypsin-like serine proteinase. *Nature* **369**, 72-76.

Almazan F, Gonzalez JM, Penzes Z, Izeta A, Calvo E, Plana-Duran J & Enjuanes L (2000). Engineering the largest RNA virus genome as an infectious bacterial artificial chromosome. *Proc. Natl. Acad. Sci. USA* **97**, 5516-5521.

Anand K, Palm GJ, Mesters JR, Siddell G, Ziebuhr J & Hilgenfeld R (2002a). Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. *EMBO J* **21**, 3213-3224.

Anand K, Pal D & Hilgenfeld R (2002b). An overview on 2-methyl-2,4-pentanediol in crystallization and in crystals of biological macromolecules. *Acta Crystallogr.* D**58**, 1722-1728

Andino R, Rieckhof GE, Achacoso PL & Baltimore D (1993). Poliovirus RNA synthesis utilizes an RNP complex formed around the 5'-end of viral RNA. *EMBO J.* **12**, 3587-3598.

Asagi M, Ogawa T, Minetoma T, Sato K & Inaba Y (1986). Detection of transmissible gastroenteritis virus in feces from pigs by reversed passive hemagglutination. *Am. J. Vet. Res.* **47**, 2161-2164.

Aurora R, Sirinivasan R & Rose GD (1994). Rules for α-helix termination by glycine. *Science* **264**, 1126-1130.

Bacon DJ & Anderson WF (1988). A fast algorithm for rendering space-filling molecule pictures. *J. Mol. Graphics* **6**, 219-220.

Benbacer L, Kut E, Besnardeau L, Laude H & Delmas B (1997). Interspecies aminopeptidase-N chimeras reveal species-specific receptor recognition by canine coronavirus, feline infectious peritonitis virus, and transmissible gastroenteritis virus. *J. Virol*. **71**, 734-747.

Baric RS & Yount B (2000). Subgenomic negative-strand function during mouse hepatitis virus infection. *J. Virol*. **74**, 4039-4046.

Bazan JF & Fletterick RJ (1988). Viral cysteine protease are homologous to the trypsin-like family of serine proteases: structural and functional implications. *Proc. Natl. Acad. Sci. USA* **85**, 7872-7876.

Benbacer L, Kut E, Besnardeau L, Laude H & Delmas B (1997). Interspecies aminopeptidase-N chimeras reveal species-specific receptor recognition by canine coronavirus, feline infectious peritonitis virus, and transmissible gastroenteritis virus. *J. Virol.* **71**, 734-737.

Benfield DA, Jackwood DJ, Bac I, Saif LJ & Wesley RD (1991). Detection of transmissible gastroenteritis virus using cDNA probes. *Arch. Virol.* **116**, 91-106.

Bergmann EM, Mosimann SC, Chernaia MM, Malcolm BA & James MNG (1997). The refined crystal structure of the 3C gene product from hepatitis A virus: specific proteinase activity and RNA recognition. *J. Virol.* **71**, 2436-2448.

Blessing RH, Guo DY & Langs DA (1996). Statistical expectation value of the Debye–Waller factor and E(hkl) values for macromolecular crystals. *Acta Crystallogr.* D**52**, 257-266.

Blessing RH & Smith GD (1999). Difference structure factor normalization for heavy atom or anomalous scattering substructure determinations. *J. Appl. Crystallogr.* **32**, 664-670.

Blom N, Hansen J, Blass D & Brunak S (1996). Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Protein Sci.* **5**, 2203-2216.

Blundell TL & Johnson LN (1976). *Protein crystallography*. Academic Press, London.

Boggon TJ & Shapiro L (2000). Screening for phasing atoms in protein crystallography. *Structure Fold. Des.* **8**, R143-R149.

Brandl M, Weiss MS, Jabs A, Suhnel J & Hilgenfeld R (2001). C-H...π-interactions in proteins. *J. Mol. Biol.* **307**, 357-377.

Brian DA, Dennis DE & Guy JS (1980). Genome of porcine transmissible gastroenteritis virus. *J. Virol.* **34**, 410-415.

Brierley I, Boursnell ME, Binns MM, Bilimoria B, Blok VC, Brown TD & Inglis SC (1987). An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV. *EMBO J*. **6**, 3779-3785.

Britton P & Page KW (1990). Sequence of the S gene from a virulent British field isolate of transmissible gastroenteritis virus. *Virus Res.* **18**, 71-80.

Brünger AT, Kuriyan J & Karplus M. (1987). Crystallographic R-factor refinement by molecular dynamics. *Science* **235**, 458-60.

Brünger AT (1988). *X-PLOR Manual*, version 5.1. Yale Univ., New Haven, USA.

Brünger AT, Krukowski A & Erickson JW (1990). Slow cooling protocols for crystallographic refinement by simulated annealing. *Acta Crystallogr.* A**46**, 583-93.

Brünger AT (1992a). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472-475.

Brünger AT (1992b). *X-PLOR (v. 3.1): A system for X-ray crystallography and NMR*. Yale University Press, New Haven, Connecticut, USA.

Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T & Warren GL (1998a). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr.* D**54**, 905-921.

Brünger AT, Adams PD & Rice LM (1998b). Recent development for the efficient crystallography refinement of macromolecular structures. *Curr. Opin. Struct. Biol.* **8**, 606-611.

Callebaut P, Pensaert MB & Hooyberghs J (1989). A competitive inhibition ELISA for the differentiation of serum antibodies from pigs infected with transmissible gastroenteritis virus (TGEV) or with the TGEV-related porcine respiratory coronavirus. *Vet. Microbiol*. **20**, 9-19.

Carugo O & Argos P (1997). Protein-protein crystal-packing contacts. *Protein Sci.* **6**, 2261-2263.

Carretero J, Sanchez F, Montero M, Blanco E, Riesco JM, Carbajo E, Gonzalez R & Vazquez R (1990). Morphological and functional study of the GH-immunoreactive adenohypophyseal cells in ovariectomized rats. *Histochem. J.* **22**, 683-687.

Caspar LD & Badger J (1991). Plasticity of crystalline proteins. *Curr. Biol.* **1**, 877-882.

Cavanagh D (1997). *Nidovirales*: a new order comprising *Coronaviridae* and *Arteriviridae*. *Arch. Virol.* **142**, 629-633.

Cavanagh D & Horzinek MC (1993). Genus torovirus assigned to the *Coronaviridae*. *Arch. Virol.* **128,** 395-396.

Cavanagh D (1995). The coronavirus surface glycoprotein. In Siddell SG (ed.), *The Coronaviridae*, Plenum Press, New York and London, pp. 73-113.

Cavanagh D, Brian DA, Enjuanes L, Holmes KV, Lai MM, Laude H, Siddell SG, Spaan W, Taguchi F & Talbot PJ (1990). Recommendations of the coronavirus study group for the nomenclature of the structural proteins, mRNAs, and genes of coronaviruses. *Virology* **176**, 306-307.

Chothia C (1973). Conformation of twisted β-pleated sheets in proteins. *J. Mol. Biol.* **75**, 295-302.

Chou PY & Fasman GD (1978). Empirical predictions of protein conformation. *Annu. Rev. Biochem*. **47**, 251-276.

Chou KC (1997). Prediction and classification of α–turn types. *Biopolymers* **42**, 837-853.

Cohen GE (1997). ALIGN: a program to superimpose protein coordinates, accounting for insertions and deletions. *J. Appl. Crystallogr.* **30**, 1160-1161.

Collaborative Computational Project Number 4 (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* D**50**, 760-763.

Cowtan KD & Main P (1996). Phase combination and cross validation in iterated density-modification calculations. *Acta Crystallogr.* D**52**, 43-48.

Cowtan KD (1999). Error estimation and bias correction in phase-improvement calculations. *Acta Crystallogr.* D**55**, 1555-1567.

Creighton (1993). *Proteins – structure and molecular properties*. Freeman WH & Company, New York.

David-Ferreira JF & Manaker RA (1965). An electron microscope study of the development of a mouse hepatitis virus in tissue culture cells. *J. Cell. Biol.* **24**, 57-78.

De La Fortelle E & Bricogne G (1997). Maximum-likelihood heavy atom parameter refinement in the MIR and MAD methods. *Methods Enzymol*. **276**, 472-494.

De Vries AAF, Horzinek MC, Rottier JM & de Groot RI (1997). The genome organization of the *Nidovirales*: similarities and differences between arteriviruses, toroviruses and coronaviruses. *Semin. Virol*. **8,** 33-47.

Deacon AM & Ealick SE (1999). Se-based MAD phasing: setting the sites on larger structures. *Structure* **7**, R161-R166.

DeLano WL (2002). The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos CA, USA. http://www.pymol.org/

Delmas B, Gelfi J, L'Haridon R, Vogel LK, Sjostrom H, Noren O & Laude H (1992). Aminopeptidase N is a major receptor for the entero-pathogenic coronavirus TGEV. *Nature* **357**, 417-420.

den Boon JA, Snijder EJ, Chirnside ED, de Vries AA, Horzinek MC & Spaan WJ (1991). Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily. *J. Virol.* **65**, 2910-2920.

Diederichs K & Karplus PA (1997). Improved R-factor for diffraction data analysis in macromolecular crystallography. *Nature Struct. Biol*. **5**, 269-275.

Dougheri AM, Storz J, Hajer I & Fernando HS (1976). Morphology and morphogenesis of a coronavirus infecting intestinal epithelial cells of newborn calves. *Exp. Mol. Pathol*. **25**, 355-370.

Dougherty WG & Semler BL (1993). Expression of virus-encoded proteinases: functional and structural similarities with cellular enzymes. *Microbiol. Rev.* **57**, 781-822.

Eleouet JF, Rasschaert D, Lambert P, Levy L, Vende P & Laude H (1995). Complete sequence (20 kilobases) of the polyprotein-encoding gene 1 of transmissible gastroenteritis virus. *Virology* **206**, 817-822.

Enjuanes L, Smerdou C, Castilla J, Anton IM, Torres JM, Sola I, Golvano J, Sanchez JM, Pintado B (1995). Development of protection against coronavirus induced diseases. A review. *Adv. Exp. Med. Biol.* **380**, 197-211.

Enjuanes L & van der Zeijst BAM (1995). Molecular basis of transmissible gastroenteritis virus epidemiology. In Siddell,S.G. (ed.). *The Coronaviridae*. Plenum Press, New York, NY, pp. 337-376.

Esnouf R.M. (1999). Further additions to Molscript version 1.4, including reading and contouring electron density maps. *Acta Crystallogr.* D**55**, 938-940.

Fujinaga M, Delbaere LTJ, Brayer GD & James MNG (1985). Refined structure of α-lytic protease at 1.7Å resolution. *J. Mol. Biol.* **184**, 479-502.

Fujinaga M, Sielecki AR, Read R, Ardelt W, Laskowski Jr. M & James MNG (1987). Crystal and molecular structures of the complex of α-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 Å resolution. *J. Mol. Biol.* **195**, 397-418.

Garwes DJ (1988). Transmissible gastroenteritis. *Vet. Rec.* **122**, 462-463.

Gewirth D (1997). The HKL manual: *A description of the programs DENZO, XDISPLAYF, and SCALEPACK: An oscillation data processing suite for macromolecular crystallography*.

Gilbert D, Westhead D, Nagano N & Thornton JM (1999). Motif-based searching in TOPS protein topology databases. *Bioinformatics* **15**, 317-326.

Godet M, L'Haridon R, Vautherot JF, Laude H (1992). TGEV corona virus ORF4 encodes a membrane protein that is incorporated into virions. *Virology* **188**, 666-675.

Godet M, Grosclaude J, Delmas B & Laude H (1994). Major receptor-binding and neutralization determinants are located within the same domain of the transmissible gastroenteritis virus (coronavirus) spike protein. *J. Virol.* **68**, 8008-8016.

Gorbalenya AE, Koonin EV, Donchenko AP & Blinov VM (1989a). Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis. *Nucl. Acids Res.* **17**, 4847-4861.

Gorbalenya AE, Donchenko AP, Blinov VM & Koonin EV (1989b). Cysteine protease of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold. *FEBS Lett.* **243**, 103-114.

Gorbalenya AE, Wassenaar AL & Snijder EJ (1989c). Arterivirus $N_{SP}2$ cysteine endopeptidase. Berrett AJ, Rawlings ND & Woessner JF (eds.). *The handbook of Proteolytic Enzymes*. Academic press, San Diego , USA pp. 693-95.

Guex N and Peitsch MC (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.

Guo DY, Blessing RH & Hauptman HA (1991). On integrating the techniques of direct methods with anomalous dispersion. II. Statistical properties of the two-phase structure invariants. *Acta Crystallogr.* A**47**, 340-345.

Hämmerle T, Molla A & Wimmer E (1992). Mutational analysis of the proposed FG loop of poliovirus proteinase 3C identifies amino acids that are neccessary for 3CD cleavage and might be determinants of a function distinct from proteolytic activity. *J. Virol*. **66**, 6028-6034.

Harris KS, Xiang W, Alexander L. Lane WS, Paul AV & Wimmer E (1994). Interaction of poliovirus polypeptide $3CD^{pro}$ with the 5' and 3' termini of the poliovirus genome. *J. Biol. Chem.* **269**, 27004-27014.

Hegyi A, Friebe A, Gorbalenya AE & Ziebuhr J (2002). Mutational analysis of the active centre of coronavirus 3C-like proteases. *J. Gen. Virol.* **83,** 581-593.

Hegyi A & Ziebuhr J (2002). Conservation of substrate specificities among coronavirus main proteases. *J. Gen. Virol.* **83,** 595-599.

Hendrickson WA, Horton JR & LeMaster DM (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J*. **9**, 1665-1672.

Hendrickson WA (1991). Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51-58.

Hendrickson WA & Ogata CM (1997). Phase determination from multiwavelength anomalous diffraction measurements. *Methods Enzymol.* **276,** 494-523.

Hendrickson WA, Smith JL, Phizackerley RP & Merritt EA (1988). Crystallographic structure analysis of lamprey hemoglobin from anomalous dispersion of synchrotron radiation. *Proteins* **4**, 77-88.

Herold J, Raabe T, Schelle-Prinz B & Siddell SG (1993). Nucleotide sequence of the human coronavirus 229E RNA polymerase locus. *Virology* **195**, 680-691.

Herold J, Siddell SG & Ziebuhr J (1996).Characterization of coronavirus RNA polymerase gene products. *Methods Enzymol*. **275**, 68-89.

Herold J, Gorbalenya AE, Thiel V, Schelle B & Siddell SG (1998). Proteolytic processing at the amino terminus of human coronavirus 229E gene 1-encoded polyproteins: identification of a papain-like proteinase and its substrate. *J. Virol*. **72**, 910-918.

Herron JN, Terry AH, Johnston S, He X, Guddat LW, Vose Jr. EW & Edmundson AB (1994). High resolution structures of the 4-4-20 Fab-fluorescein complex in two solvent systems; effects of solvent on structure and antigen-binding affinity. *Biophys J.* **67**, 2167-2183.

Holm L & Sander (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.

James MNG, Sielecki AR, Brayer GD, Delbaere LTJ & Bauer CA (1980). Structure of product and inhibitor complexes of *Streptomyces griseus* protease A at 1.8 Å resolution: a model for serine protease catalysis. *J. Mol. Biol.* **144**, 43-88.

Janin J & Rodier F (1995). Protein-protein interaction at crystal contacts. *Proteins* **23**, 580-587.

Jaskolski M & Wlodawer A (1996). A minimalist approach to the phase-problem – phasing selenomethionyl protein structures using Cu K_α data. *Acta Crystallogr.* D**52**, 1075-1081.

Johnston S & Holgate S (1996). Epidemiology of viral respiratory tract infections. In Myint S & Taylor-Robinson D (eds.), *Viral and other infections of the human respiratory tract.* Chapman and Hall, London, United Kingdom, pp. 1-38.

Jones TA, Cowan S, Zou J-Y & Kjeldgaard M (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr.* A**47**, 110-119.

Jones TA & Kjeldgaard M (1995). O-the manual (Version 5.11). Upspsala Univ., Uppsala, Sweden.

Kabsch W & Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.

Kamphuis IG, Kalk KH, Swarte MB & Drenth J (1984). Structure of papain refined at 1.65 Å resolution. *J. Mol. Biol.* **179**, 233-256.

Kapke PA, Tung FY, Hogue BG, Brian DA, Woods RD & Wesley R (1988). The amino-terminal signal peptide on the porcine transmissible gastroenteritis coronavirus matrix protein is not an absolute requirement for membrane translocation and glycosylation. *Virology* **165**, 367-376.

Karplus PA & Faerman C (1994). Ordered water in macromolecular structures. *Curr. Opin. Struct. Biol.* **4**, 770-778.

Khan AR, Khazanovich-Bernstein N, Bergmann EM & James MN (1999). Structural aspects of activation pathways of aspartic protease zymogens and viral 3C protease precursors. *Proc. Natl. Acad. Sci. USA* **96**, 10968-10975.

Kleywegt GJ & Jones TA (1997). Good model-building and refinement practice. *Methods Enzymol*. **277**, 208-230.

Kleywegt GJ, & Jones TA (1998). Databases in protein crystallography. *Acta Crystallogr.* D**54**, 1119-1131.

Krantz A, Copp LJ, Coles PJ, Smith RA & Heard SB (1991). Peptidyl (acyloxy)methyl ketones and the quiescent affinity label concept: the departing group as a variable structural element in the design of inactivators of cysteine proteinases. *Biochemistry* **30,** 4678-4687.

Kraulis PJ (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946-950.

Kräusslich HG & Wimmer E (1988). Viral proteinases. *Annu. Rev. Biochem.* **57**, 701-754.

Kubo H, Yamada YK & Taguchi F (1994). Localization of neutralizing epitopes and the receptor-binding site within the amino-terminal 330 amino acids of the murine coronavirus spike protein. *J. Virol*. **68**, 5403-5410.

Kuo L, Godeke G-J, Raamsman JBM, Masters PS & Rottier PJM. **(**2000). Retargeting of coronavirus by substitution of the spike glycoprotein ectodomain: crossing the host cell species barrier. *J. Virol.* **74,** 1393-1406.

Kyte J & Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **15**, 105-132.

Lai M & Cavanagh D (1997). The molecular biology of coronaviruses. *Adv. Virus Res*. **48,** 1-100.

Lai MM, Baric RS, Brayton PR & Stohlman SA (1984). Characterization of leader RNA sequences on the virion and mRNAs of mouse hepatitis virus, a cytoplasmic RNA virus. *Proc. Natl. Acad. Sci. USA* **81**, 3626-3630.

Lämmli U (1970). Cleavage of structural proteins during the assembly of the head of bacterophage T4. *Nature* **227**, 680-685.

Lamzin VS & Wilson KS (1997). Automated refinement for protein crystallography. *Methods. Enzymol*. **277**, 269-305.

Laskowski RA, MacArthur MW, Moss DS & Thornton JM (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26, 283-291.

Lee B & Richards FM (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol*. **55**, 379-400.

Lee HJ, Shieh CK, Gorbalenya AE, Koonin EV, La Monica N, Tuler J, Bagdzhadzhyan A & Lai MM (1991). The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative protease and RNA polymerase. *Virology* **180**, 567-582.

Leong LE, Walker PA & Porter AG (1993). Human rhinovirus-14 protease 3C ($3C^{pro}$) binds specifically to the 5'-noncoding region of the viral RNA. *J. Biol. Chem.* **268**, 25735-25739.

Liu DX & Brown TD (1995). Characterisation and mutational analysis of an ORF 1a-encoding proteinase domain responsible for proteolytic processing of the infectious bronchitis virus 1a/1b polyprotein. *Virology* 209, 420-427.

Liu DX, Xu HY & Brown TD (1997). Proteolytic processing of the coronavirus infectious bronchitis virus 1a polyprotein: identification of a 10-kilodalton polypeptide and determination of its cleavage sites. *J. Virol.* **71**, 1814-1820.

Love RA, PargeHE, Wickersham JA, Hostomsky Z, Habuka N, Moomaw EW, Adachi ?? & Hostomska Z. (1996). The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site. *Cell* **87**, 331-342.

Lu G (1999). FINDNCS: A program to detect non-crystallographic symmetries in protein crystals from heavy atom sites. *J. Appl. Crystallogr.* **32**, 365.

Lu Y & Denison MR (1997). Determinants of mouse hepatitis virus 3C-like proteinase activity. *Virology* **230**, 335-342.

Lu Y, Lu X & Denison MR (1995). Identification and characterization of a serine-like proteinase of the murine coronavirus MHV-A59. *J. Virol.* **69**, 3554-3559.

Lu X, Sims AC & Denison MR (1998). Mouse hepatitis virus 3C-like protease cleaves a 22-kilodalton protein from the open reading frame 1a polyprotein in virus-infected cells and in vitro. *J. Virol.* 72, 2265-2271.

Luzzati V (1952). Traitement statistique des errors dans la determination des structures cristallines. *Acta Crystallogr.* **5**, 802-810.

MacArthur MW, Laskowski RA & Thornton JM (1994). Validation of protein models derived from experiment. *Curr. Opin. Struct. Biol.* **4**, 731-737.

Malcolm BA (1995). The picornaviral 3C proteinases: cysteine nucleophiles in serine proteinase folds. *Protein Sci.* **4**, 1439-1445.

Martin Alonso JM, Balbin M, Garwes DJ, Enjuanes L, Gascon S & Parra F (1992). Antigenic structure of transmissible gastroenteritis virus nucleoprotein. *Virology* **188**, 168-174.

Matthews BW (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491-497.

Matthews DA, Smith WW, Ferre RA, Condon B, Budahazi G, Sisson W, Villafranca JE, Janson CA, McElroy HE, Gribskov CL & Worland S (1994). Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site, and means for cleaving precursor polyprotein. *Cell* **77**, 761-771.

Matthews DA, Dragovich PS, Webber SE, Fuhrman SA, Patick AK, Zalman LS, Hendrickson TF, Love RA, Prins TJ, Marakovits JT, Zhou R, Tikhe J, Ford CE, Meador JW, Ferre RA, Brown EL, Binford SL, Brothers MA, DeLisle DM & Worland ST (1999). Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. *Proc. Natl. Acad. Sci. USA* **96**, 11000-11007.

McGregor MJ, Islam SA & Sternberg MJE (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295-310.

McDonald I & Thornton JM (1994). Satisfying hydrogen bonding potentials in proteins. *J. Mol. Biol.* **238**, 777-793.

McPherson A (1998). *Crystallization of biological macromolecules*. Cold Spring Harbour Laboratory Press, Cold Spring Harbor, New York.

Menard R, Plouffe C, Laflamme P, Vernet T, Tessier DC, Thomas DY & Storer AC (1995). Modification of the electrostatic environment is tolerated in the oxyanion hole of the cysteine protease papain. *Biochemistry* **34**, 464-471.

Merritt EA & Bacon DJ (1997). Raster3D: photorealistic molecular graphics. *Meth. Enzymol*. **277**, 505-524.

Miller R, Gallo S, Khalak HG & Weeks CM (1994). *SnB*: crystal structure determination via Shake-&-Bake. *J. Appl. Crystallogr.* **27**, 613-621.

Moradian-Oldak J, Leung W & Fincham AG (1998). Temperature and pH-dependent supramolecular self-assembly of amelogenin molecules: a dynamic light-scattering analysis. *J. Struct. Biol*. **122**, 320-327.

Mosimann SC, Cherney MM, Sia S, Plotch S & James MN (1997). Refined X-ray crystallographic structure of the poliovirus 3C gene product. *J. Mol. Biol.* **273**, 1032-1047.

Myint SH (1995). Human coronavirus infections. In Siddell SG (ed.). *The Coronavirdae*. Plenum Press, New York, pp. 389-401.

Navaza J (1994). AMoRe: An automated package for molecular replacement. *Acta Crystallogr.* A**50**, 157-163.

Nataraj AJ, Trent JC 2nd & Ananthaswamy HN (1995). p53 gene mutations and photocarcinogenesis. *Photochem Photobiol.* 62, 218-230.

Ng LF & Liu DX (2000). Further characterization of the coronavirus infectious bronchitis virus 3C-like proteinase and determination of a new cleavage site. *Virology* **272**, 27-39.

Nicholls A, Sharp KA, & Honig B (1991). Protein folding and association: insights from the interfacial and thermodynamic properties. *Proteins* **11**, 281-296.

Otwinowski Z (1991). Maximum-likelihood refinement of heavy atom parameters. In: Wolf W, Evans PR & Leslie AGW (eds.), *Proceedings of the CCP4 Study Weekend: Isomorphous Replacement and Anomalous Scattering.* SERC Proceedings, Daresbury Laboratories, Warrington, UK, pp. 80-88.

Otwinoski Z & Minor W (1997). Processing of X-ray diffraction data collected in oscillation mode. *Meth. Enzymol* **276**, 307-326.

Palmenberg AC (1990). Proteolytic processing of picornaviral polyprotein. *Annu. Rev. Microbiol.* **44,** 603-623.

Pannu NS, Murshudov GN, Dodson EJ & Read RJ (1998). Incorporation of prior phase information strengthens maximum-likelihood structure refinement. *Acta Crystallogr.* D**54**, 1285-1294.

Parks GD, Baker JC & Palmenberg AC (1989). Proteolytic cleavage of encephalomyocarditis virus capsid region substrates by precursors to the 3C enzyme. *J. Virol.* **63**, 1054-1058.

Pavone V, Gaeta G, Lombardi A, Nastri F & Maglio O (1996). Discovering protein secondary structures: classification and description of isolated alpha-turns. *Biopolymers* **38**, 705-721.

Penzes Z, González JM, Calvo E, Izeta A, Smerdou C, Méndez A, Sánchez CM, Sola I, Almazán F & Enjuanes L (2001). Complete genome sequence of transmissible gastroenteritis coronavirus PUR46-MAD clone and evolution of the purdue virus cluster. *Virus Genes* **23,** 105-118.

Perrakis A, Morris R, Lamzin VS (1999). Automated protein model building combined with iterative structure refinement. Nat Struct.l Biol. **6**, 458-463.

Petersen JF, Cherney MM, Liebig HD, Skern T, Kuechler E & James MN (1999). The structure of the 2A proteinase from a common cold virus: a proteinase responsible for the shut-off of host-cell protein synthesis. *EMBO J.* **18**, 5463-5475.

Pinon JD, Teng H & Weiss SR (1997). Further requirements for cleavage by the murine coronavirus 3C-like proteinase: identification of a cleavage site within ORF1b. *Virology* **263**, 471-484.

Plagemann PG & Moennig V (1992). Lactate dehydrogenase-elevating virus, equine arteritis virus, and simian hemorrhagic fever virus: a new group of positive-strand RNA viruses. *Adv. Virus Res.* **41**, 99-192.

Polgár L (1974). Mercaptide-imidazolium ion-pair: the reactive nucleophile in papain catalysis. *FEBS Lett*. **47**, 15-18.

Qian C, Lagace L, Massariol M-J, Chabot C, Yoakim C, Deziel R & Tong L (2000). A rational approach towards successful crystallization and crystal treatment of human cytomegalovirus protease and its inhibitor complex. D**56**, 175-180.

Ramachandran GN & Sasisekharan V (1968). Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23**, 283-438.

Rashin AA, Iofin M & Honig B (1986). Internal cavities and buried waters in globular proteins. *Biochemistry* **25**, 3619-3625.

Ravelli RBG, Sweet RM, Skinner JM, Duisenberg AJM & Kroon J (1997). STRATEGY: a program to optimize the starting angle and scan range for X-ray data collection. *J. Appl. Crystallogr.* **30**, 551-554.

Read RJ (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr.* A**42**, 140-9.

Read RJ (1996). As MAD as can be. *Structure* **4**, 11-14.

Regula G, Lichtensteiger CA, Mateus-Pinilla NE, Scherba G, Miller GY, Weigel RM (2000). Comparison of serologic testing and slaughter evaluation for assessing the effects of subclinical infection on growth in pigs. *J. Am. Vet. Med. Assoc*. **217**, 888-895.

Richards FM (1977). Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151-176.

Richardson JS (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem*. **34**, 167-339.

Richardson JS, & Richardson DC (1988). Amino acid preferences for specific locations at the ends of α helices. *Science* **240**, 1648.

Richardson JS, Getzoff ED & Richardson DC (1978). The β bulge: a common small unit of nonrepetitive protein structure. *Proc. Natl. Acad. Sci. USA* **75**, 2574-2578.

Rossmann MG (1990). The molecular replacement method. *Acta Crystallogr.* A**46**, 73-82.

Rossmann MG & Blow DM (1962). The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr.* **15**, 24-31.

Rost B & Sander C (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol*. **232**, 584-599.

Russell RB, Saqi MA, Bates PA, Sayle RA & Sternberg MJ (1998). Recognition of analogous and homologous fold protein folds-assessment of prediction success and associated alignment accuracy using empirical substitutions matrices. *Protein Eng*. **11**, 1-9.

Rupley JA. Yang PH & Tollin G (1980). In water, Rowland SP (ed.). *In polymers* Amer. Chem. Soc., Washington DC, pp. 111-120.

Saif LJ & Wesley R (1999). Transmissible gastroenteritis virus. In Straw BES, Allaire WL, Mengeling WL & Taylor DJ (eds.), *Diseases of Swine* (8th Edition), Iowa State University Press, Ames, Iowa, pp. 295-325.

Sanchez A, Ossorio C, Alvaro-Gracia JM, Padilla R & Avila J (1990). A subset of antibodies from the sera of patients with systemic lupus erythematosus react with vimentin and DNA. *J. Rheumatol*. **17**, 205-209.

Satow Y, Cohen GH, Padlan EA & Davies DR (1986). Phosphocholine binding immunoglobulin Fab McPC603. An X-ray diffraction study at 2.7 Å. *J. Mol. Biol.* **190**, 593-604. (ALIGN)

Schaad MC & Baric RS (1994). Genetics of mouse hepatitis virus transcription: evidence that subgenomic negative strands are functional templates. *J. Virol*. **68,** 8169-8179.

Schellman JA (1980). The α_L conformation at the ends of helices. In Jaenicke R (ed.), *Protein folding*, Elsevier, Amsterdam, North Holland, pp. 53-61.

Schneider-Schaulies J (2000). Cellular receptors for viruses: links to tropism and pathogenesis. *J. Gen. Virol.* (2000), **81**, 1413-1429.

Schoenborn BP (1965). Binding of xenon to horse haemoglobin. *Nature* **208**, 760-762.

Schiller JJ, Kanjanahaluethai A & Baker SC (1998). Processing of the coronavirus MHV-JHM polymerase polyprotein: identification of precursors and proteolytic products spanning 400 kilodaltons of ORF1a. *Virology* **242**, 288-302.

Schultze B, Gross HJ, Brossmer R, Klenk HD & Herrler G (1990). Hemagglutinating encephalomyelitis virus attaches to N-acetyl-9-O-acetylneuraminic acid-containing receptors on erythrocytes: comparison with bovine coronavirus and influenza C virus. *Virus Res.* **16**, 185-194.

Sethna PB, Hofmann MA & Brian DA (1991). Minus-strand copies of replicating coronavirus mRNAs contain antileaders. *J. Virol.* **65,** 320-325.

Sethna PB, Hung SL & Brian DA (1989). Coronavirus subgenomic minus-strand RNAs and the potential for mRNA replicons. *Proc. Natl. Acad. Sci. USA* **86,** 5626-5630

Seybert A, Hegyi A, Siddell SG, & Ziebuhr J (2000). The human coronavirus 229E superfamily 1 helicase has RNA and DNA duplex-unwinding activities with 5'-to-3' polarity. *RNA* **6,** 1056-1068.

Seybert A, van Dinten LC, Snijder EJ, & Ziebuhr J (2000). Biochemical characterization of the equine arteritis virus helicase suggests a close functional relationship between arterivirus and coronavirus helicases. *J. Virol.* **74**, 9586-9593.

Seybert A, Ziebuhr J & Siddell SG (1997). Expression and characterization of a recombinant murine coronavirus 3C-like proteinase. *J. Gen. Virol.* **78**, 71-75.

Sheldrick GM (1990a). Phase annealing in *SHELX*-90: direct methods for larger structures. *Acta Cryst.* A**46**, 467-473.

Sheldrick GM (1990b). Computing aspects of crystal structure determination. *J. Mol. Struct*. **130**, 9-16.

Sheldrick GM (1991). Tutorial on automated Patterson methods to find heavy atoms. In Moras D, Podjarny AD & Thierry JC (eds.), *Crystallographic computing 5*, pp. 145-157.

Sibanda BL & Thornton JM (1985). Beta-hairpin families in globular proteins. *Nature* **316**, 170-174.

Siddell SG (1995). *The Coronaviridae: An introduction*. Plenum Press, New York, pp. 1-10.

Siddell S, Sawicki D, Meyer Y, Thiel V & Sawicki S (2001). Identification of the mutations responsible for the phenotype of three MHV RNA-negative Ts mutants. *Adv. Exp. Med. Biol*. **494**, 453-458.

Simkins RA, Weilnau PA, Bias J & Saif LJ (1992). Antigenic variation among transmissible gastroenteritis virus (TGEV) and porcine respiratory coronavirus strains detected with monoclonal antibodies to the S protein of TGEV. *Am. J. Vet. Res.* **53**, 1253-1258.

Snijder EJ & Spaan WJ (1995). The coronavirus-like superfamily. Siddell SG (ed.). *The Coronavirdae.* Plenum press, London & New York, pp. 239-255.

Spaan W, Delius H, Skinner M, Armstrong J, Rottier P, Smeekens S, van der Zeijst BA & Siddell SG (1983). Coronavirus mRNA synthesis involves fusion of non-contiguous sequences. *EMBO J.* **2**, 1839-1844.

Sternberg MJE & Thornton JM (1977). On the conformation of proteins: hydrophobic ordering of strands in β-pleated sheets. *J. Mol. Biol.* **115**, 1-17.

Storer A & Menard R (1994). Catalytic mechanism in papain family of cysteine peptidases. *Methods Enzymol*. **244**, 486-500.

Suck D (1997). Common fold, common function, common origin? *Nature Struct. Biol.* **4**, 161-165.

Sugiyama K & Amano Y (1980). Hemagglutination and structural polypeptides of a new coronavirus associated with diarrhea in infant mice. *Arch. Virol.* **66**, 95-105.

Terwilliger TC & Berendzen J (1996). Bayesian weighting for macromolecular crystallographic refinement. *Acta Crystallogr.* D**52**, 743-748.

Thiel V, Herold J, Schelle B & Siddell SG (2001). Infectious RNA transcribed *in vitro* from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. *J. Gen. Virol.* **82**, 1273-81.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F & Higgins DG (1997). The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876-4882.

Tibbles KW, Brierly I, Cavanagh D & Brown TD (1996). Characterization in vitro of an autocatalytic processing activity associated with the predicted 3C-like proteinase domain of the coronavirus avian infectious bronchitis virus. *J. Virol.* **70**, 1923-1930.

Tibbles KW, Cavanagh D & Brown TD (1999). Activity of a purified His-tagged 3C-like proteinase from the coronavirus infectious bronchitis virus. *Virus Res.* **60**, 137-45.

Tilton RF, Kintz ID & Petsko GA (1984). Cavities in proteins: structure of a metmyoglobin-xenon complex solved to 1.9 Å. *Biochemistry* **23**, 2849-2857.

Tilton RF, Singh PC, Weiner SJ, Connolly ML, Kuntz ID, Kollman PA, Max N & Case DA (1986). Computational studies of the interaction of myoglobin and xenon. *J. Mol. Biol.* **192**, 443-456.

Tooze J, Tooze S & Warren G (1984). Site of addition of N-acetyl-galactosamine to the E1 glycoprotein of mouse hepatitis virus-A59. *Eur J. Cell. Biol*. **33**, 281-293.

Tsukada H and Blow DM (1985). Structure of α-chymotrypsin refined at 1.68 Å resolution. *J. Mol. Biol.* **184**, 703-711.

Uson I, Pohl E, Schneider TR, Dauter Z, Schmidt A, Fritz HJ & Sheldrick GM (1999). 1.7 Å Structure of the stabilised REIv mutant T39K. Application of local NCS restraints. *Acta Crystallogr.* D**55**, 1158-1167.

Vaguine AA, Richelle J & Wodak SJ (1999). SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr.* D**55**, 191-205.

van der Meer Y, Snijder EJ, Dobbe JC, Schleich S, Denison MR, Spaan WJ & Locker JK (1999). Localization of mouse hepatitis virus nonstructural proteins and RNA synthesis indicates a role for late endosomes in viral replication. *J Virol.* **73**, 7641-7657.

van Dinten LC, Rensen S, Gorbalenya AE & Snijder EJ (1999). Proteolytic processing of the open reading frame 1b-encoded part of arterivirus replicase is mediated by nsp4 serine protease and is essential for virus replication. *J. Virol*. **73**, 2027-2037.

van Marle G, van Dinten LC, Spaan WJ, Luytjes W & Snijder EJ (1999). Characterization of an equine arteritis virus replicase mutant defective in subgenomic mRNA synthesis. *J. Virol*. **73**, 5274-5281.

Venkatachalam CM (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**, 1425-1436.

Vriend G (1990). WHATIF: A molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52-56.

Watson HC (1969). *Prog. Stereochem*. **4**, 299.

Weeks CM and Miller R (1999). The design and implementation of SnB version 2.0. *J. Appl. Crystallogr.* **32**, 120-124.

Weingartl HM & Derbyshire JB (1994). Evidence for a putative second receptor for porcine transmissible gastroenteritis virus on the villous enterocytes of newborn pigs. *J. Virol*. **68**, 7253-7259.

Weiss MS & Hilgenfeld R (1997). On the use of merging R-factor as a quality indicator for X-ray data. *J. Appl. Crystallogr.* **30**, 203-205.

Weiss MS, Palm GJ & Hilgenfeld R (2000). Crystallization, structure solution and refinement of hen egg-white lysozyme at pH 8.0 in the presence of MPD. *Acta Crystallogr.* D**56**, 952-958.

Weiss MS (2001). Global indicators of X-ray data quality. *J. Appl. Crystallogr.* **34**, 130-135.

Wilmot CM & Thornton JM (1988). Analysis and prediction of the different types of β-turn in proteins. *J. Mol. Biol.* **203**, 221-232.

Xiang W, Harris KS, Alexander L & Wimmer E (1995). Interaction between the 5'-terminal cloverleaf and 3AB/3CDpro of poliovirus is essential for RNA replication. *J. Virol.* **69**, 3658-3667.

Xu H, Hauptman HA, Weeks CM & Miller R (2000). P1 Shake-&-Bake: can success be guaranteed? *Acta Crystallogr.* D**56**, 238-240.

Xu H, Weeks CM, Deacon AM, Miller R & Hauptman HA (2000). III-conditioned Shake-&-Bake: the trap of the false solution. *Acta Crystallogr.* A**56**, 112-118.

Yao Z, Jones DH & Grose C (1992). Site-directed mutagenesis of herpesvirus glycoprotein phosphorylation sites by recombination polymerase chain reaction. *PCR Methods Appl.* **1,** 205-207.

Yeager CL, Ashmun RA, Williams RK, Cardellichio CB, Shapiro LH, Look AT & Holmes KV (1992). Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature* **357**, 420-422.

Yu X, Bi W, Weiss SR & Leibowitz JL (1994). Mouse hepatitis virus gene 5b protein is a new virion envelope protein. *Virology* **202**, 1018-1023.

Zhang XM, Herbst W, Konsoulas KG & Storz J (1994). Biological and genetic characterization of hemagglutinating coronavirus isolated from a diarrhoeic child. *J. Med. Virol.* **44**, 152-161.

Zhang X & Matthews BW (1994). Conservation of solvent-binding sites in 10 crystal forms of T4 lysozyme. *Protein Sci.* **3**, 1031-1039.

Ziebuhr J, Herold J & Siddell SG (1995). Characterization of a human coronavirus (strain 229E) 3C-like proteinase activity. *J. Virol.* **69**, 4331-4338.

Ziebuhr J, Heusipp G, Seybert A & Siddell SG (1997a). Substrate specificity of the human coronavirus 229E 3C-like proteinase. *Adv. Experimental Med. Biol.* **440**, 115-120.

Ziebuhr J, Heusipp G & Siddell SG (1997b). Biosynthesis purification, and characterization of the human coronavirus 229E 3C-like proteinase. *J. Virol.* **71**, 3992-3997.

Ziebuhr J & Siddell SG (1999). Processing of the human coronavirus 229E replicase polyproteins by the virus-encoded 3C-like proteinase: identification of proteolytic products and cleavage sites common to pp1a and pp1ab. *J. Virol.* **73**, 177-185.

Ziebuhr J, Snijder EJ & Gorbalenya AE (2000). Virus-encoded proteinases and proteolytic processing in the *Nidovirales*. *J. Virol.* **81**, 853-879.

6. APPENDIX

6.1 Crystal parameters and refinement statistics

Table 6.1A Crystal parameters and statistics of diffraction data for TGEV M^{pro} – CMK (substrate analog chloromethyl ketone) inhibitor complex

X-ray source	Synchrotron radiation[a]
Space group	$P2_1$
Unit cell (Å, °)	a = 72.39, b = 158.55, c = 88.20, β = 94.40
Resolution (Å)	50-2.37 (2.41-2.37)
Total no. of reflections	562107
No. of reflections used	562107
Unique reflections	78630
Completeness (%)	99.0
Average intensity I/σ(I)	9.9 (3.4)
R_{merge}[b] (%)	5.8
R_{pim}[c] (%)	2.2

[a] X-ray diffraction at Joint IMB Jena-University of Hamburg-EMBL synchrotron beamline X13 at Deutsches Elektronen-Synchrotron, Hamburg, equipped with a Mar CCD detector
[b,c] For definition see Section 2.2.12

Table 6.1B Refinement and model statistics of TGEV M^{pro} – CMK substrate analog complex

No. of non-hydrogen atoms (average B value (Å^2))	
Protein	13911 (43.0)
Water	925 (51.3)
MPD	32 (78.6)
Sulfate	135 (59.8)
R-factor[a]	19.2
Free R-factor	23.6
Rmsd from ideal geometry	
Bonds (Å)	0.006
Bond angles (°)	1.3

[a, b] For definition see Section 2.2.12

6.2 Inter-subunit H-bonds

Table 6.2.1 Inter-subunit (intra-dimer) H-bond contacts (< 3.5 Å) for all monomers: Main chain–main chain*

Residue involved				Distance (Å)
Phe 139A	O	Ser 1B	O	2.88
Phe 139A	N	Ser 1B	O	3.49
Ala 7B	N	Val 124A	O	3.20
Ala 7B	O	Val 124A	N	2.85*
Val 124B	N	Ala 7A	O	2.78
Val 124B	O	Ala 7A	N	3.21
Phe 139B	O	Ser 1A	N	2.88
Ser 279B	O	Gly 281A	N	3.05
Phe 139C	O	Ser 1D	N	2.59
Phe 139C	N	Ser 1D	O	3.32
Ala 7D	N	Val 124C	O	3.21
Ala 7D	O	Val 124C	N	3.00*
Val 124D	N	Ala 7C	O	2.79
Val 124D	O	Ala 7C	N	3.27
Phe 139D	O	Ser 1C	N	2.59
Ser 279D	O	Gly 281C	N	2.82
Gly 281D	N	Ser 279C	O	3.08*
Ala 7E	N	Val 124F	O	3.38
Ala 7E	O	Val 124F	N	2.72*
Val 124E	N	Ala 7F	O	2.71
Val 124E	O	Ala 7F	N	3.13
Phe 139E	O	Ser 1F	N	3.47
Ser 279E	O	Gly 281F	N	3.48
Gly 281E	N	Ser 279F	O	3.49*
Phe 139F	O	Ser 1E	N	3.47

*All these H-bond are present in HCoV M^{pro} (AB dimer) as well, with exceptions of *

Table 6.2.2 Inter-subunit (intra-dimer) H-bond contacts (< 3.5 Å): Main chain–side chain*

Residue name				Distance (Å)
Ser 1A	N	Glu 165B	OE1	2.77
Arg 4A	NH1	Gly 126B	O	2.98
Arg 4A	NH1	Lys 136B	O	3.21
Arg 4A	**NH2**	**Gly 126B**	**O**	2.60
Gly 11A	N	Glu 14B	OE2	2.81
Gly 126A	O	Arg 4B	NH1	2.94
Arg 130A	NH2	Glu 193A	OE2	2.99
Gly 126A	**O**	**Arg 4B**	**NH2**	3.20
Phe 139A	O	Ser 1B	OG	2.97
Arg 4B	NH1	Gly 126A	O	2.94
Arg 4B	**NH2**	**Gly 126A**	**O**	3.20
Gly 11B	**N**	**Glu 14A**	**OE2**	2.59
Glu 14B	**OE2**	**Gly 11A**	**N**	2.81
Gly 126B	O	Arg 4A	NH1	2.98
Gly 126B	**O**	**Arg 4A**	**NH2**	2.60
Lys§ 136B	**O**	**Arg 4A**	**NH1**	3.21
Ser 138B	OG	Gly 2A	O	3.25
Arg 4C	**NH2**	**Gly 126D**	**O**	3.21
Arg 4C	N	Gln 295D	OE1	2.70
Gly 11C	**N**	**Glu 14D**	**OE2**	2.67
Glu 14C	**OE2**	**Gly 11D**	**N**	2.72
Lys§ 136C	**O**	**Arg 4D**	**NH2**	2.41
Ser 1D	OG	Gly 137C	O	2.53
Gly 11D	**N**	**Glu 14C**	**OE2**	2.72
Glu 14D	**OE1**	**Gly 11C**	**N**	2.67
Gly 126D	**O**	**Arg 4C**	**NH2**	3.21
Ser 138D	OG	Gly 2C	O	3.09
Glu 165D	**OE1**	**Ser 1C**	**N**	3.08
Arg 4E	NH1	Gly 126F	O	2.87
Arg 4E	**NH2**	**Gly 126F**	**O**	2.57
Glu 14E	OE2	Gly 11F	N	2.85
Gly 136E	O	Arg 4F	NH1	3.23
Ser 136E	O	Arg 4F	NH2	3.28
Glu 139E	O	Ser 1F	N	3.47
Glu 165E	**OE1**	**Ser 1F**	**OG**	3.28
Glu 165E	**OE2**	**Ser 1F**	**OG**	3.39
Arg 4F	NH1	Lys 136E	O	3.23
Arg 4F	**O**	**Gln 295F**	**NE2**	2.82
Gly 11F	N	Glu 14E	OE2	2.85
Glu 126F	O	Arg 4E	CZ	2.80
Gly 126F	O	Arg 4E	NH1	2.87
Ser 126F	O	Arg 4E	NH2	2.57
Ser 138F	OG	Met 6E	CE	3.09

* Interactions satisfying H-bond critria in HCoV M^{pro} (AB dimer) are marked in bold

**An extra interaction in HCoV M^{pro}: is SerA 10 N - GluB14 OE1 2.70 Å

§ replaced by Arg in HCoV M^{pro}

Table 6.2.3 Inter-subunit (intra-dimer) H-bond contacts (< 3.5 Å): Side chain–side chain*

Residues involved				Distance (Å)
Ser 1A	**OG**	**Glu 165B**	**OE1**	2.77
Ser 1A	**OG**	**Glu 165B**	**OE2**	2.86
Arg 40A	**NH2**	**Asp 186A**	**OD2**	2.89
Arg 40A	**NH2**	**Asp 186A**	**OD1**	3.47
Arg 40A	**NE**	**Asp 186A**	**OD1**	2.44
Lys 103A	NZ	Glu 152A	OE2	2.87
Glu 165A	OE1	Ser 1B	OG	3.50
Arg 40B	**NH2**	**Asp 186B**	**OD2**	2.92
Arg 40B	**NE**	**Asp 186B**	**OD1**	2.57
Asp 34B	OD2	Lys 89B	NZ	3.13
Arg 130B	NH2	Glu 193B	OE1	2.92
Ser 138A	OG	Arg 4B	CG	3.45
Ser 138B	OG	Gln 295A	OE1	3.07
Glu 152B	OE2	Lys 103B	NZ	3.05
Glu 165B	OE1	Ser 1A	OG	3.08
Glu 165B	OE1	His 171B	NE2	2.72
Glu 165B	OE2	Ser 1A	OG	2.86
Glu 14C	OE2	Ser 10 D	OG	2.81
Arg 40C	**NE**	**Asp 186C**	**OD1**	2.60
Arg 40C	**NH2**	**Asp 186C**	**OD2**	3.05
Arg 40C	**NH2**	**Asp 186C**	**OD1**	3.35
Arg 40C	**NH2**	**Arg 4D**	**NH2**	3.22
Arg 130C	NH2	Glu 193C	OE1	3.02
Ser 138C	OG	Arg 4D	NE	3.20
Ser 138C	OG	Asp 186C	OD2	3.05
Glu 152C	OE2	Lys 103C	NZ	3.05
Glu 165C	OE1	His 171C	NE2	3.16
Arg 4D	**NE**	**Ser 138C**	**OG**	3.20
Arg 4D	**NH2**	**Ser 138C**	**OG**	3.22
Arg 40D	NE	Asp 186D	OD1	2.61
Arg 40D	NH2	Asp 186D	OD2	2.98
Arg 40D	NH2	Asp 186D	OD1	3.22
Gly 11D	N	Glu 14C	OE2	2.72
Ser 138D	OG	Gln 295C	OE1	2.85
Glu 165D	OE2	Ser 1C	OG	3.34
Arg 40E	**NE**	**Asp 186E**	**OD1**	2.63
Arg 40E	**NH2**	**Asp 186E**	**OD2**	3.26
Lys 103E	NZ	Glu 152B	OE2	3.40
Arg 130E	NE	Glu 193E	OE1	3.21
Glu 152E	OE2	Lys 103E	NZ	3.40
Glu 165E	OE1	Ser 1F	OG	3.28
Glu 165E	OE2	Ser 1F	OG	3.39
Ser 1F	**OG**	**Glu 165E**	**OE1**	**3.39**
Arg 40F	**NE**	**Asp 186F**	**OD1**	2.69

Arg 40F	**NH2**	**Asp 186F**	**OD2**	2.74
Arg 40F	**NH2**	**Asp 186F**	**OD1**	3.45
Ser 138F	OG	Gln 295F	OE1	3.35
Glu 152F	OE2	Lys 103F	NZ	3.12
Glu 165 F	OE1	His 1 71F	NE2	2.95

* Interactions satisfying H-bond critria in HCoV M^{pro} (AB dimer) are marked in bold

**Three extra interactions in HCoV M^{pro}: Glu286 OE2 - Arg4 NH2 2.85Å
Lys294[§] NZ - Gln122[#] OE1 3.29 Å
Ser138 OG –Arg4 N 3.50 Å

[§] replaced by Arg in TGEV M^{pro}
[#] replaced by Gly in TGEV M^{pro}

Table 6.2.4 Inter-dimer H-bond contacts (< 3.5 Å):

Residues involved				Distance (Å)
Lys 101D	NZ	Asp 245B	OD1	3.03
Glu 185D	OE1	Val 50E	N	2.87
Gln 191 D	NE2	Arg 49E	NH1	3.26
Lys 234E	NZ	179B	OE2	2.82
Arg 216E	NH1	Arg 275B	NH2	2.46
Val 219E	O	Gly 274B	N	2.88
Lys 234E	NZ	Glu 179B	OE2	2.82
Gly 273E	O	Arg 275A	NH2	2.82
Arg 275E	NH1	Gly 273A	O	2.99
Asn 55F	ND2	Gly 182C	O	2.96
Ser 58F	OG	Gln 132C	OE1	2.61
Ser 78F	OG	Lys 234C	O	2.96
Arg 80F	NE	Lys 234C	O	3.13
Arg 80F	NH2	Phe 238C	O	2.87
Lys 101F	NZ	Asp 245A	OD2	3.90

Table 6.3A Buried water molecules in the M^{pro} crystal structure having accessible surface area (ASA) <1.0 $Å^{2§}$

Water No.	**B-factor**	**H-bonding partners**			
TGEV M^{pro} Dimer AB	**($Å^2$)**		**TGEV M^{pro} Dimer CD**	**TGEV M^{pro} Dimer EF**	**HCoV M^{pro}**
W 2	30.1	202 O, 284 O, 286 N, 287 N	13	658	
W 3	30.5	202 O, 284 O, 286 N, 287 O, 287 N	57	14	
W 4	29.4	5 O, 126 N	63	53	208
W 7	22.4	126 N, 5 O	70	75	
W 11	33.1	41 ND1, 163 ND1, 40 N	258	321	48
W 15	39.8	41 N, 41 ND1, 163 ND1, 186 OD1 40 N	204	28	128
W 23	36.5	19 NH1, 19 O, 66 OG, 66 O, 68 N, 68 OG	197	56	
W 41	34.7	102 O, 102 N	264	565	
W 47	34.7	110 O	327	687	
W 53	33.1	5 O, 126 N	758	493	

W 54	34.4	133 O, 173 O, 173 N, 181 O	795	8	
W 58	44.4	131 N, 132 N, 193 OE1, 196 OD1		1003	
W 59	44.5	128 OD1, 129 O, 285 OD1	158	925	
W 77	38.4	9 O, 12 N, 11 N	381	76	
W 100	34.0	9 O, 10 O, 11 N, 12 N	46	42	45
W105	37.9	19 O, 19 NH1, 66 O, 68 N, 68 OG	423	313	
W 193	33.5	108 O, 202 OD1	33	60	
W 217	44.1	128 OD1, 285 O, 285 OD1, 285 OD2	243	-	
W 231	39.1	108 O, 201 OD1, 201 OD2, 202 OD1		150	
W 310	44.7	128 O, 136 N, 286 OE1	122	754	
W 549	47.4	45 N, 46 N, 46 O, 56 OE1	-	367	
W 551	57.6	14 OE1	389	412	
W 570	35.7	14 O, 17 O, NZ 69	191	69	
W580	53.5	40 NH2, 185 OE2, 186 OD2, NE2 187	266	696	
W 611	38.1	14 OE1	830	648	153
W704	40.3	132 O, 182 O, 184 O	538	113	28
W 772	49.9	138 O, 138 OG	308	997	
W 880	34.8	4 NH2, 5 O	290	709	
W944	22.9	124A O, 124B O	234	156	
W 964	41.9	4 NH1, 5 O	249	386	

§ Buried in both structures; Buried only in one structure; highlighted are the conserved water molecules in the active-site cavity

Table 6.3B Buried water molecules in TGEV M^{pro} structure but not conserved in all TGEV M^{pro} dimers:

W 22	21.9	103 O, 103 N, 158 O, 176 N			
W 25	31.6	131 N, 196 OD1			
W 49	35.2	131 N, 197 O, 196 ND2			
W 55	30.1	222 O, 220 N, 260 O, 220 OG1			
W117	40.0	103 O, 103 N, 158 O, 176 N			5
W120	43.3	154 O, 156 N			
W121	41.1	128 N, 136 O, 286 OE1 OE2			
W122	44.5	128 O, 286 OE1			
W 198	37.9	107 N, 130 O, 131 O			
W 222	54.0	279 O, 281 N, 282 N			
W 223	45.7	169 O, 136 N			
W 292	50.6	108 N, 129 O, 285 OD1			
W 293	40.3	107 O			
W 313	52.4	19 NH1, 19 O, 66 O, 68 N, 68 OG			
W 321	40.6	40 N, 41 N, 41 ND1, 163 ND1, 186 OD1			
W 322	43.9	87 O, 80 N, 78 OG			
W 336	43.3	272 O			
W 345	42.7	199 OG			

W 347	52.4	104 O, 109 OE2, 109 OE1			
W359	41.8	175 ND2, 179 O			168
W367	48.8	45 N, 46 O, 45 OG, 46 N			
W384	46.4	169 O, 1 OG			
W 386	46.0	5 O			
W 411	40.9	45 N, 45 OG, 46 O, 46 N, 56 OE1			
W 416	46.7	197 O			
W 418	42.8	220 N, 222 O, 267 NH2, 260 O			
W 459	39.5	40NH1, 83 O			
W467	48.1	108 N, 107 O			
W 477	41.0	133 O, 173 N, 181 O			
W 569	36.6	131 N, 196 OD1, O 197			
W 587	48.6	193 O, 195 O, 195 N			
W 612	53.1	103 O, 158 O, 176 N			
W 658	54.5	284 N, 285, 286, 287, 202 O			
W 724	49.2	4 NH2, 136 O, 286 OE2			
W 753	54.2	130 O, 131 O, 133 N			
W 902	48.6	119 O, 69 NZ,			
W 966	52.3	295 NE2, 291 O			
W 990	40.0	162 NE2			38
W997	55.3	138 O, 138 OG			

§ Buried in both structures; Buried only in one structure; highlighted are the conserved water molecules in the active-site cavity

Table 6.3C Buried water molecules in the HCoV M^{pro} crystal structure

W5	1.75	103 O, 158 O, 176 N
W14	4.53	132 O, 194 N, W100
W16	5.06	202 O, 206 N
W21	6.29	130 O, 131 N, 131 OG1, 196 ND2
W25	7.24	41 N, 163 OE1, 186 OD1
W32	9.06	140 O, 143 N, W19
W39	10.24	W49, 138 O, W145
W45	11.60	9 O, 10 O, 10 N, 11 N, 12 N,
W53	12.41	202 O, 286 N, 287 N
W93	18.36	6 N, 291 OE2
W128	21.97	40 N, 41 N, 41 ND1, 84 O,
W152	25.19	138 O, 117 OH, W9
W168	27.78	W162
W170	27.98	145 O
W215	49.22	175 OG, 179 O

§ Buried in both structures; Buried only in one structure

Table 6.3D Spatially conserved exposed water molecules in the HCoV and TGEV M^{pro} crystal structures having accessible surface area (ASA) >4.0 $Å^2$

Water No.		**Water No.**	
HCoV M^{pro}	**HCoV M^{pro} interactions**	**TGEV M^{pro}**	**TGEVM^{pro} interactions**
W6…….	W125, 178 O	W88…….	178 O
W7…….	102 N	W642……..	102 N
W12……	W151, 180 N, 27 OH	W24…….	22 O
W20……	276 OE1	W904……..	SO_4 O2, O3
W28……	182 O	W704……..	182 O, 184 OH
W31……	180 N	W749……..	180 O
W41……	184 O, W125	W242……..	185 OE2, 186 N
W51……	103 NE, 109 OE1	W79…….	104 O, 109 NZ, 109 OE1
W62……	103 O, 104 N	W181……..	103 N
W63……	15 NZ	W168……..	11 O, 14 OE1
W144……..	W87	W567……..	153 N, 154 N
W161……..	94 ND2	W408……..	15 O, 94 OD1, 96 ND2
W191.....	186 OD2, 186 N	W134……..	186 N, 186 OD2,

Table 6.4 List of salt bridges

Residue name	**Number**	**Atom name**	**Residue name**	**Number**	**Atom name**	**Distance (Å)***
TGEV M^{pro}						
Ser	**1A**	**N**	**Glu**	**165B**	**CD**	**3.68**
Arg	**4A**	**NH1**	**Glu**	**286B**	**OE1**	**5.31**
Lys	5A	NZ	Glu	291A	OE1	2.90
Glu	35A	OE2	Lys	89A	NZ	3.22
Arg	**40A**	**NE**	**Asp**	**186A**	**OD1**	2.44
Lys	103A	NZ	Glu	152A	OE1	2.71
Arg	130A	NH2	Glu	193A	OE1	2.73
Glu	193A	OE1	Arg	130A	NH2	2.73
Lys	234A	NZ	Glu	240A	OE2	2.54
Ser	**1B**	**N**	**Glu**	**165A**	**CD**	**4.50**
Arg	4B	NH1	Glu	286A	OE1	3.75
Asp	34B	OD1	Lys	89B	NZ	2.76
Arg	**40B**	**NE**	**Asp**	**186B**	**OD1**	**2.70**
Lys	103B	NZ	Glu	152B	OE1	2.76
Arg	130B	NE	Glu	193B	OE1	3.21
Glu	152B	OE1	Lys	103B	NZ	2.76
Asp	263B	OD2	Arg	267B	NH2	2.84
Glu	291B	OE1	Lys	5B	NZ	2.74
Ser	**1C**	**OG**	**Glu**	**165D**	**OE2**	3.34
Arg	4C	NH1	Glu	286D	OE1	3.45

Asp	34C	OD1	Lys	89C	NZ	3.14
Arg	**40C**	**NE**	**Asp**	**186C**	**OD1**	2.60
Arg	130C	NE	Glu	193C	OE1	2.89
Glu	152C	OE1	Lys	103C	NZ	2.43
Asp	245C	OD1	Arg	61B	NH1	2.52
Glu	291C	OE1	Lys	5C	NZ	2.59
Ser	**1D**	**N**	**Glu**	**165C**	**CD**	**6.15**
Arg	4D	NH2	Glu	286C	OE2	4.77
Arg	**40D**	**NE**	**Asp**	**186D**	**OD1**	2.61
Lys	101D	NZ	Asp	245B	OD1	3.03
Arg	130D	NE	Glu	193D	OE1	2.89
Ser	**1E**	**N**	**Glu**	**165F**	**CD**	**5.88**
Arg	4E	NH1	Glu	286F	OE1	4.67
Lys	5E	NZ	Glu	291E	OE1	2.61
Asp	34E	OD2	Lys	89E	NZ	3.33
Glu	35E	OE2	Lys	89E	NZ	3.20
Arg	**40E**	**NE**	**Asp**	**186E**	**OD1**	2.63
Lys	103E	NZ	Glu	152B	OE1	2.55
Arg	130E	NH1	Glu	193E	OE1	2.82
Glu	152E	OE1	Lys	103E	NZ	2.55
Ser	**1F**	**OG**	**Glu**	**165E**	**OE2**	**3.28**
Arg	4F	NH1	Glu	286E	OE1	5.03
Glu	35F	OE2	Lys	89F	NZ	3.29
Arg	**40F**	**NE**	**Asp**	**186F**	**OD1**	2.69
Lys	101F	NZ	Asp	245A	OD2	3.90
Arg	130F	NE	Glu	193F	OE1	3.51
Glu	152F	OE1	Lys	103F	NZ	2.90

***Only the shortest distances between any oxygen and any nitrogen atom of a given salt bridge is listed**

HCoV M^{pro}						
Lys	5A	NZ	Glu	286A	OE2	3.58
Arg	19A	NE	Asp	118A	OD2	3.43
Arg	19A	NH2	Asp	118B	OD1	3.47
Arg	**40A**	**NE**	**Asp**	**186A**	**OD1**	**2.82**
Arg	**40A**	**NH2**	**Asp**	**186A**	**OD2**	**3.24**
Arg	**40A**	**NH2**	**Asp**	**186A**	**OD1**	**3.49**
Arg	61A	NH1	His	63A	NE2	3.76
Arg	61A	NH1	His	63A	ND2	3.79
Arg	103A	NH1	Glu	157A	OE2	2.90
Arg	61A	NH2	His	63A	OE1	3.46

Asp	118A	OD2	Arg	19A	NE	3.32
Asp	118A	OD1	Arg	19A	NH2	3.28
Glu	165A	OE2	His	171A	NE2	3.60
Glu	222A	OE1	His	263A	NE2	3.03
Glu	**286A**	**OE2**	**Arg**	**4B**	**NH2**	**2.99**
Arg	4B	NH2	Glu	286B	OE2	2.99
Arg	19B	NE	Asp	118B	OD1	3.27
Arg	19B	NH2	Asp	118B	OD2	3.27
Arg	**40B**	**NE**	**Asp**	**186B**	**OD1**	**2.82**
Arg	**40B**	**NH2**	**Asp**	**186B**	**OD2**	**3.24**
Arg	**40B**	**NH2**	**Asp**	**186B**	**OD1**	**3.49**
Arg	61B	NH1	His	63B	NE2	3.67
Arg	61B	NH1	His	63B	ND2	3.70
Arg	103B	NH1	Glu	157A	OE1	3.46
Arg	103B	NH2	Glu	157A	OE2	2.90
Asp	118B	OD2	Arg	19B	NE	3.17
Asp	118A	OD1	Arg	19A	NH2	3.10
Glu	165B	OE2	His	171B	NE2	2.75
Glu	222B	OE1	His	263B	NE2	3.07
Glu	**286B**	**OE2**	**Arg**	**4A**	**NH2**	**2.85**

*Conserved salt bridges are indicated in bold

Table 6.5 Interactions with sulfates, MPD and dioxane molecules in TGEV M^{pro} structure

SO4	1S	O3	Glu	172A	OE2, OE1	3.40, 2.77
SO4	2S	S	Arg	130C	NH2	3.72
SO4	2S	O1	Arg	130C	NH2, NH1	2.79, 3.79
SO4	2S	O3	Arg	130C	NH2, NH1	3.67, 3.04
SO4	3S	S	Glu	193A	N	3.64
SO4	3S	O1	Asn	168A	ND2	3.22
SO4	3S	O3	Wat, Glu, Asn	160W, 193A, 168 A	OH2, N, ND2	2.75, 2.89, 3.83
SO4	3S	O4	Glu, Gly	193A, 194 A	N, N	3.28, 3.12
SO4	4S	S	Arg, Wat	130A, 760W	NH1, OH2	2.90, 3.70
SO4	4S	O1	Arg, Asn	130A, 196A	NH1, ND2	2.92, 3.41
SO4	4S	O2	Wat, Wat	904W, 760W	OH2, OH2	3.34, 3.82
SO4	4S	O3	Arg, Wat, Lys, Wat	130A, 904W, 136A, 760W	NH1, OH2, NZ, OH2	2.61, 3.73, 3.82, 2.79
SO4	4S	O4	Arg, Lys, Wat	130A, 136A, 760W	NH1, NZ, OH2	2.88, 3.62, 3.89
SO4	5S	S	Asn, Glu, Gly	168C, 193C, 194C	ND2, N, N	3.69, 3.80, 3.84
SO4	5S	O1	Asn, Wat	168C, 350W	ND2, OH2	2.99, 3.23

SO4	5S	O2	Glu, Gly, Wat	193C, 194C, 350W	N, N, OH2	3.57, 2.87, 3.81
SO4	5S	O3	Glu	193A	N	2.95
SO4	5S	O4	Wat, Asn, Glu	164W, 168C, 193C	OH2, ND2, N	2.75, 3.21, 2.84
SO4	6S	S	Asn, Wat, Glu	168B, 691W, 193B	ND2, OH2, N	3.60, 3.87, 3.80
SO4	6S	O2	Asn, Wat, Glu	168B, 303W, 193B	ND2, OH2, N	3.12, 2.91, 2.95
SO4	6S	O3	Asn, Wat	168B, 691W	ND2, OH2	2.95, 2.99
SO4	6S	O4	Gly, Wat, Glu	194B, 691W,193B	N, OH2, N	3.01, 3.63, 3.40
SO4	7S	S	Glu, Asn	193E, 168E	N, ND2	3.75, 3.45
SO4	7S	O2	Glu, Wat, Asn	193E, 361W, 168E	N, OH2, ND2	2.94, 3.27, 3.11
SO4	7S	O3	Wat, Asn	641W, 168E	OH2, ND2	3.32, 2.68
SO4	7S	O4	Glu, Gly, Wat	193E, 194E,641W	N, N, OH2	3.41, 3.22, 3.68
SO4	8S	S	Arg, Wat	130D, 438W	NH2, OH2	3.62, 3.65
SO4	8S	O2	Wat	438W	OH2	2.62
SO4	8S	O3	Arg, Arg	130D, 130D	NH2, NH1	3.88,3.25
SO4	8S	O4	Arg, Asn, Wat, Arg	130D, 196D, 438W, 130D	NH2, ND2, OH2, NH1	2.47, 3.21, 3.66, 3.48
SO4	9S	S	Wat, Arg,Wat, Wat	774W, 130E, 392W, 888W	OH2, NH2, OH2, OH2	3.34, 3.76, 3.50, 3.66
SO4	9S	O1	Wat, Wat, Arg, Wat, Wat	774W, 778W, 130E, 329W, 888W	OH2, OH2, NH1, OH2, OH2	2.72, 3.62, 3.59, 2.34, 3.38
SO4	9S	O2	Wat, Wat	392W, 888W	OH2, OH2	3.62, 2.78
SO4	9S	O3	Wat, Wat	426W, 774W	OH2, OH2	3.47, 2.80
SO4	9S	O4	Wat, Asn, Arg, Arg	426W, 196E, 130E, 130E	OH2, OD1, Nh2, NH2	3.80, 3.73, 3.76, 2.66
SO4	10S	S	Asn, Glu, Gly	168F, 193F, 194F	ND2, N, N	3.46, 3.54, 3.67
SO4	10S	O1	Asn, Wat, Glu, Gly	168F, 951W, 193F, 194F	ND2, OH2, N, N	2.69, 3.32, 2.45, 3.68
SO4	10S	O2	Asn, Glu, Gly	168F, 193F, 194F	NH2, N, N	3.59, 3.81, 3.10
SO4	10S	O3	Asn	168F	ND2	3.59
SO4	10S	O4	Glu, Gly	193F, 194F	N, N	3.83, 3.60
SO4	11S	S	Arg, Trp	216A, 217A	N, N	3.68, 3.72
SO4	11S	O1	Arg, Trp	216A, 217A	N, N	3.83, 2.89
SO4	11S	O2	Wat, Arg, Trp	679W, 216A, 217A	OH2, N, N	3.72, 3.11, 3.44
SO4	11S	O4	Arg, Gly	216A, 214A	N, O	3.55, 3.48
SO4	12S	S	Asn, Glu	168D,193D	ND2, N	3.73, 3.71
SO4	12S	O1	Asn,	168D	ND2	2.99
SO4	12S	O3	Asn, Glu, Wat	168D, 193D,138W	ND2, N, OH2	3.38, 2.95, 2.89
SO4	12S	O4	Glu, Gly	193D, 194D	N, N	3.43, 3.15
SO4	13S	S	Thr, Arg, Arg, Wat	276E, 275E, 275E, 349W	N, NE, NH2, OH2	3.67, 3.73, 3.77, 3.82
SO4	13S	O1	Arg	275E	NE	3.75
SO4	13S	O2	Arg, Wat	275E, 349W	NH2, OH2	3.14, 3.84
SO4	13S	O3	Thr, Thr, Arg, Arg, Wat, Thr	276E, 276E, 275E, 275E, 349E, 276E	N, E, NE, NH2, OH2, OG1	2.85, 3.29, 2.92, 3.32, 2.96, 3.69

SO4	13S	O4	Thr, Thr	276E, 276E	N, OG1	3.68, 2.77
SO4	15S	S	Thr, Thr	276C, 276C	N, OG1	3.64, 3.72
SO4	15S	O1	Arg, Thr, Thr, Thr, Wat, Arg	275C, 276C, 276C, 276C, 450W, 275C	NE, N, OG1, O, OH2, NH2	3.32, 2.76, 3.40, 3.49, 3.44, 3.45
SO4	15S	O2	Arg, Arg	275C, 275C	NE, NH2	3.52, 3.86
SO4	15S	O4	Thr, Thr	276C, 276C	N, OG1	3.36, 3.18
SO4	16S	S	Arg, Trp	216D, 217D	N, N	3.71, 3.74
SO4	16S	O1	Wat	999W	OH2	2.73
SO4	16S	O2	Arg, Arg, Trp	216D, 275D, 217D	N, NH2, N	3.45, 3.72, 3.34
SO4	16S	O3	Gly	214D	O	3.41
SO4	16S	O4	Arg, Trp	217D	N	2.99
SO4	17S	S	Thr, Thr	276D, 276D	OG1, N	3.73, 3.70
SO4	17S	O2	Arg, Arg	275D, 275D	NH2, NE	3.72, 3.55
SO4	17S	O3	Thr, Thr	276D, 276D	OG1, N	3.07, 3.27
SO4	17S	O4	Arg, Thr, Arg, Thr, Thr	275D, 276D, 275D, 276D, 276D	NH2, OG1, NE, N, O	3.50, 3.65, 3.14, 2.94, 3.36
SO4	18S	S	Leu, His	62E, 63E	N, ND1	3.66, 3.72
SO4	18S	O2	Val, Wat, Leu	60E, 699W, 62E	O, OH2, N	3.57, 3.89, 2.83
SO4	18S	O3	Leu, His, His	62E, 63E, 63E	N, N ND1	3.60, 3.84, 2.66
SO4	18S	O4	Val, Arg, His	60E, 61E, 63E	O, NH2, ND1	3.68, 2.66, 3.70
SO4	19S	S	Dox, Arg	3X, 216F	O1, NH1	3.56, 3.21
SO4	19S	O2	Arg	216F	NH1	3.47
SO4	19S	O3	Dox, Arg	3X, 216F	O1, NH1	3.49, 2.82
SO4	19S	O4	Dox, Arg	3X, 216F	NH1	3.02, 2.86
SO4	20S	S	Phe	176B	O	3.89
SO4	20S	O3	Wat	919W	OH2	3.89
SO4	20S	O4	Phe	176B	O	2.77
SO4	21S	O2	Arg	294C	O	3.65
SO4	21S	O3	Ala	115D	O	3.59
SO4	21S	O4	Ser	123D	OG	3.64
SO4	22S	S	Ser	104A	N	3.63
SO4	22S	O3	Ser	104A	N	3.15
SO4	22S	O4	Ser, Wat, Ser	104A, 181W, 104A	OG, OH2, N	3.22, 3.82, 3.00
SO4	23S	S	Arg, Trp	216C, 217C	N, N	3.73, 3.74
SO4	23S	O1	Arg, Trp	216C, 217C	N, N	2.42, 2.50
SO4	23S	O2	Arg, Wat, Arg	275C, 366W, 275C	NH1, OH2, NH2	3.75, 2.80, 3.90
SO4	23S	O3	Gly	214C	O	3.64
SO4	23S	O4	Wat, Arg	450W, 275C	OH2, NH2	3.23, 3.66
SO4	24S	S	Leu, Arg	62F, 61F	N, NH2	3.77, 3.49
SO4	24S	O1	Leu, His, Arg, Arg, Arg	62F, 61F, 62F, 61F, 61F	N, ND1, NE,NH1, NH2	3.49, 3.06, 3.75, 3.74, 3.20
SO4	24S	O3	Leu, Arg	62F, 61F	N, NH1	3.14, 3.85

SO4	24S	O4	Arg, Arg	61F, 61F	NH1, NH2	3.61, 2.68
SO4	25S	S	Arg	130F	NH2	3.14
SO4	25S	O1	Arg	130F	NH2	2.84
SO4	25S	O2	Arg	130F	NH2	3.43
SO4	25S	O4	Lys, Arg	136F, 130F	NZ, NH2	3.18
SO4	26S	S	Arg, Wat, Ser	19B, 288W, 21B	NH2, OH2, OG	3.47, 3.55, 3.71
SO4	26S	O1	Arg, Wat, Arg, Ser, Wat	19B, 288W, 19B, 21B, 494W	NH2, OH2, NH1, OG, OH2	2.94, 3.16, 2.62, 3.01, 2.93
SO4	26S	O2	Arg, Wat	19B, 288W	NH2, OH2	3.80, 2.77
SO4	26S	O3	Wat	486W	OH2	3.84
SO4	26S	O4	Arg, Ser	19B, 21B	NH2, OG	3.19, 3.59
SO4	27S	S	Ser, Arg	21A, 19A	OG, NH2	3.39, 3.64
SO4	27S	O1	Ser	21A	OG	3.39
SO4	27S	O2	Ser, Arg	21A, 19A	OG, NH2	3.87, 2.99
SO4	27S	O3	Ser, Arg, Arg, Ser	21A, 19A, 19A, 66A	OG, NH2, NH1, OG	3.64, 3.01
MPD	1M	O4	His	41A, 163A	NE2, O	3.36, 2.92
MPD	2M	O2	Thr, Pro	47B, 188B	O, N	3.18, 3.85
MPD	2M	O4	Thr, Asp	47B, 186B	O, O	3.76, 3.79
MPD	3M	O2	Wat	681	OH2	3.18
MPD	4M	O2	Gln, Pro	187D, 188D	O, N	3.90, 3.47
MPD	4M	O4	Wat	640	OH2	3.50
MPD	5M	O2	Thr	47E	O	3.72
MPD	5M	O4	His	41E	CG	3.55
MPD	6M	O4	His	41F	ND1	3.55
DOX	1X	O1	Gly, Wat, Wat, Wat	273B, 1006W, 199W, 1005W	N, OH2, OH2, OH2	3.54, 2.97, 3.76, 3.51
DOX	1X	O2	Trp, Arg, Wat, Wat	217E, 216E, 524W, 199W	N, N, OH2, OH2	3.50, 3.57, 2.89, 2.86
DOX	3X	O1	SO4, SO4	19S, 19S	S, O4	3.56, 3.02
DOX	3X	O2	Gly, Thr, Thr	244F, 276F, 276F	O, N, OG1	3.55, 2.92, 3.58
DOX	4X	O1	Asn, Met	112A, 6A	OD1, O	3.61, 3.76
DOX	4X	O2	Wat	926W	OH2	3.67
DOX	5X	O1	Gly, Ser, Arg	133A, 131A, 130A	N, O, O,	3.46, 2.72, 3.82
DOX	6X	O2	Wat	620W	OH2	3.13
DOX	7X	O1	Wat	898W	OH2	3.70
DOX	7X	O2	Phe, His	139B, 162B	O, NE2	3.48, 3.83
DOX	8X	O1	Ser	223C	O	3.08
DOX	8X	O2	Ser	228C	OG	2.70
DOX	9X	O1	Asn	213D	OD1	2.99
DOX	9X	O2	Wat, Lys, Wat	461W, 136C, 998W	OH2, NZ, OH2,	3.45, 2.73, 3.82

Selbständigkeitserklärung

Ich erkläre, daß ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegeben Hilfsmittel und Literatur angefertigt habe.

Jena, im Dezember 2002 (Kanchan Anand)

Erklärung zur Bewerbung

Ich erkläre, daß ich mich mit der vorliegenden Arbeit an keiner anderen Hochschule um den akademischen Grad doctor rerum naturalium beworben habe und daß ich weder früher noch gegenwärtig die Eröffnung eines Verfahrens zum Erwerb des oben genannten akademischen Grades an einer anderen Hochschule beantragt habe.

Jena, im Dezember 2002 (Kanchan Anand)